PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G01N 33/00, G06F 17/00 — A1

(11) International Publication Number: **WO 98/48270**

(43) International Publication Date: 29 October 1998 (29.10.98)

(21) International Application Number: PCT/US98/08077

(22) International Filing Date: 21 April 1998 (21.04.98)

(30) Priority Data:

60/044,124	22 April 1997 (22.04.97)	US
Not furnished	20 April 1998 (20.04.98)	US

(71) Applicant: CALIFORNIA INSTITUTE OF TECHNOLOGY [US/US]; 1200 East California Boulevard, Pasadena, CA 91125 (US).

(72) Inventors: GODDARD, William, A., III; 1200 East California Boulevard, Pasadena, CA 91125 (US). DEBE, Derek, A.; 1200 East California Boulevard, Pasadena, CA 91125 (US).

(74) Agents: RING, Christine, S. et al.; Limbach & Limbach L.L.P., 2001 Ferry Building, San Francisco, CA 94111–4262 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD OF DETERMINING THREE–DIMENSIONAL PROTEIN STRUCTURE FROM PRIMARY PROTEIN SEQUENCE

(57) Abstract

The Generic Protein method is a computer–implemented system for determining the three–dimensional structure of a protein from its amino acid sequence. The method incorporates a hierarchical approach wherein the number of candidate structures decreases at each step. The starting point is the use of a sequence independent ensemble of compact structures which represents an exhaustive enumeration of all possible self–avoiding folded topologies for a n residue polypeptide. Because the number of candidate conformations is dramatically reduced, recognition filters such as radius of gyration, distribution of hydrophobic residues, and the satisfaction of disulfide constraints can be used to further reduce the number of candidate conformations. The complexity of the initial ab initio structure prediction problem can be reduced to a complexity on the order of a homology modeling exercise. The final refinement step may involve molecular mechanics procedures with explicit solvation parameters on full–atom representations of the remaining candidate structures.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD OF DETERMINING THREE-DIMENSIONAL PROTEIN STRUCTURE FROM PRIMARY PROTEIN SEQUENCE

This application claims the benefit of and incorporates by reference herein U.S. Provisional Application No. 60/044,124, filed April 22, 1997 entitled "Method of Determining Three-Dimensional Protein Structure From Primary Sequence" by inventors William A. Goddard III and Derek A. Debe.

The U.S. Government has certain rights in this invention pursuant to Grant No. CHE-95-22179 and ACS-92-17368 awarded by the National Science Foundation.

BACKGROUND

The present invention generally relates to methods for determining a protein's three-dimensional structure from its amino acid sequence. More particularly, the present invention relates to methods for generating a sequence independent ensemble of folded topologies for a *n* residue protein which can then be used as a starting point in protein structure prediction. Both sequence dependent and sequence independent methods for reducing the number of potential conformations are also described.

Since the seminal work by C.B. Anfinsen, determining the three-dimensional structure of a protein from its amino acid sequence has been a much sought after goal in structural and computational biology. However, although progress has been made in several fronts such as secondary structure

prediction and homology modeling, a general method for *ab initio* structure prediction, or in other words, a solution to the so-called "protein folding problem", has eluded investigators.

In general, two reasons are most often cited for the intractability of this problem. The first relates to the combinatorics involved when attempting to exhaustively enumerate all possible conformations even when simplifying assumptions are made. For example, assuming only three conformations per amino acid residue, a relatively small protein of a 100 residue would have approximately 10^{48} (or 3^{100}) potential conformations. Since proteins in their native environments are known to fold in millisecond timescales, the apparent paradox (termed Levinthal's paradox) on how a protein arrives at its "correct" three-dimensional structure without systematically sampling the inordinately large number of potential conformations has been the subject of intense debate.

Notwithstanding the apparent paradox, various methods, such as molecular dynamics ("MD"), Monte Carlo, and Genetic Algorithm ("GA"), have been developed to sample the available conformational space. Since it was generally assumed that an exhaustive enumeration of the possible conformations was not possible, the goal of these methods have been to sufficiently sample the available conformation space so that a structure corresponding to the "native" protein may be found among the candidate structures. However, what constitutes sufficient sampling remains an open question.

The second reason for the intractability of the folding problem relates to recognition of the "native" protein structure among the candidate set. Although numerous recognition methods are described in the literature, they are not generally suitable for use with low resolution structures like those generated by *ab initio* modeling procedures.

For example, existing methods are often based on contact potentials. These methods typically use some form of a reduced representation of the protein such as an α and β carbon sphere model. In this representation, the peptide backbone is approximated by an α-carbon sphere and the amino acid sidechains are approximated by the 20 unique β-carbon spheres. The energy of the protein, E, is then approximated by the sum of all of the individual contact energies such that,

$$E_{contact} = \sum_{i} \sum_{j<i-1} \varepsilon_{\beta_i\beta_j}(R_{\beta_i\beta_j}) + \varepsilon_{\alpha_i\alpha_j}(R_{\alpha_i\alpha_j}) + \varepsilon_{\alpha_i\beta_j}(R_{\alpha_i\beta_j}) + \varepsilon_{\beta_i\alpha_j}(R_{\beta_i\alpha_j}).$$

Values of the ε terms are compiled according to the statistical relation:

$$\varepsilon_{ij} = -\ln\left(\frac{n_{ij}}{n_{ijexp}}\right),$$

where n_{ij} is the number of contacts (interactions within a predetermined cutoff distance) between interacting centers *i* and *j*, and n_{ijexp} is the number of these contacts expected in a random distribution.

Because structures generated by *ab initio* modeling procedures tend to be low resolution structures, contact potential based methods do not adequately discriminate between structures that have the correct overall fold from those that do not. Since most of these methods were developed to recognize the native crystal structure from a group of grossly misfolded decoys generated by threading the native sequence over all segments of equal length of available in the Protein Data Bank, their poor performance on lower resolution structures is not surprising.

Various attempts have been made to modify the contact potential based methods for use with *ab initio* modeling procedures. However, because local interactions are still emphasized over global protein architectures, these methods still are generally unsuitable for use with structures generated from *ab initio* modeling procedures.

Thus, an important feature of a recognition method for use in an *ab initio* context is the de-emphasis of local energies or contacts in favor of global protein architecture. The challenge presented is to accurately estimate the locally minimum energy accessible by any structure that is similar in conformation to the examined structure.

Despite the overwhelming challenges, there is an increasing need for accurate protein structure prediction methodologies as the number of known primary sequences continues to far outpace the number of solved three dimensional structures. This need is exacerbated since analytical solution of proteins of interest by either NMR or X-ray crystallography are not always possible due to experimental difficulties such as protein insolubility or insufficient quantities of purified protein. Since determining the three-dimensional structure is often the key in elucidating the function and mechanism of the protein of interest (thereby allowing researchers, for example, to discover new drugs or more potent drugs), an accurate and easy to use method for *ab initio* structure prediction would be an invaluable tool.

SUMMARY OF THE INVENTION

The present invention presents a novel approach to the protein folding problem by reframing the question in terms of folded topologies. In other words, instead of starting with the entirety of the available conformational space, the starting point for the inventive method is the subset of conformational space that represents distinct, self-avoiding, folded topologies.

By reframing the question, the number of conformations that needs to be considered is dramatically reduced from 3^n to approximately 1.2^n.

In addition, novel recognition procedures based only on the α-carbon position have also been developed. These recognition techniques analyze a protein structure to determine whether or not the hydrophobic residues are capable of accessing the protein center, and whether the charged residues are capable of accessing solvent exposed protein exterior. Because energy scores obtained by these methods are invariant over a larger CRMS range from the native structure, these methods are better suited for use with *ab initio* procedures.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic of the inventive *ab initio* protein modeling protocol wherein the three dimensional protein structure is modeled based on its amino acid (or even gene sequence).

Figure 2 is an illustration of certain variables used in one of the ensemble reduction embodiments. The residues are designated as hydrophobic and non-hydrophobic (or hydrophilic). For each hydrophobic residue, a vector is calculated from the protein center of mass to the Cα of the hydrophobic residue. For each hydrophilic residue, a cone is drawn with its vertex as the protein center of mass that encompasses the hydrophilic residue of interest.

Figure 3 is the CRMS data between the *ab initio* generated structure (GP equivalent) and the 277 native proteins and protein domains considered.

Figure 4 is a plot of the GP ensemble size required to sample each and every native folded topology as a function of polypeptide length *n*.

Figure 5 is a plot of the maximum time scale for protein folding versus the protein length based on a diffusion constant, $D=3x10^{-6}cm^2/sec$ (or one new topology approximately every 0.3 nanosecond for a 100 residue protein).

Figure 6 is a hierarchical strategy for *ab initio* protein folding. The conformation search at each step is greatly reduced due to coarse grain eliminations of conformations at the previous levels. The GP method coupled with an appropriate recognition algorithm produces a manageable set of candidates which contains the native folded topology. The time scales shown are estimates for a single processor Silicon Graphics Inc. (SGI) workstation.

Figure 7 is the comparisons of the GP structures superimposed on the corresponding native protein backbone. (a) 65-residue segment from the NMR determined structure of the proteolytic fragment from Bacteriorhodopsin (1bct): the GP structure has a CRMS fit of 5.78 Å and the refined structure has a CRMS of 4.35 Å; (b) 65 residue Porcine C5a (1c5a): the GP structure has a CRMS fit of 5.40 Å and the refined structure has a CRMS of 3.91 Å; (c) 80 residue fragment from acyl-coenzyme A binding protein (1aca): the GP structure has a CRMS fit of 6.12 Å and the refined structure has a CRMS of 4.97 Å; and, (d) 80 residue segment from domain four of the N-terminal domain of 70 kD heatshock cognate protein (1hpm04): the GP structure has a CRMS fit of 6.14 Å and the refined structure has a CRMS of 4.22 Å.

Figure 8 illustrates one implementation of a distance constrained method for generating protein structures.

Figure 9 is a flow chart of one implementation of the enrichment/replication process.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention generally relates to methods for determining a protein's three-dimensional structure from its amino acid sequence. More particularly, the present invention relates to methods for generating a sequence independent ensemble of folded topologies for a *n* residue protein which can then be used as a starting point in a hierarchical approach to protein structure prediction. Novel recognition methods which are more suitable for use with *ab initio* structure prediction procedures are also described.

The inventive methods may be implemented by being programmed into and executed on a computer and includes the three general steps. The first involves generating an ensemble of all possible tertiary folds for a *n* residue protein. The exhaustive enumeration is possible because the number of self-avoiding folded conformations is substantially smaller than all possible conformations for a *n* residue protein. Moreover, because the conformation set is only dependent on the number of residues, *n*, the initial conformation set is entirely sequence independent.

Once the initial conformation set is created, the second step involves reducing the number of potential structures by considering sequence specific information. Methods which may be used include a novel recognition protocol based only on α-carbon positions. The remaining structures may then be further refined using any number of techniques known in the art including sophisticated energy calculation (including explicit solvent) on full atom representations of the protein. Figure 1 is a schematic illustration of the procedures involved in obtaining a three-dimensional structure from a protein's primary sequence.

Each step, ensemble generation, ensemble reduction, and ensemble refinement, will be further discussed in turn.

Ensemble Generation

Although it is not necessary for the implementation of the present invention, for computational expediency, a reduced representation of the protein is generally preferred. One example of a reduced representation is the main atoms of the peptide backbone (N, Cα, C, and O). Another example is an α-carbon backbone and pseudo sidechain representation where the different amino acid sidechains are represented by a vector from the α-carbon to a pseudo Cβ position. However, because the ensemble generation is independent of the amino acid sequence of the protein, the use of a peptide backbone or an α-carbon only representation is most preferred.

The two general assumptions made in the implementation of the inventive methods are that conformations with overlapping peptide chains are inordinately high in energy and that non-neighboring spatially adjacent residues are subject to attractive van der Waals interactions. Because all amino acids are preferably treated alike at this stage, the ensemble generated for *n* residues will be the set of initial candidate structures for any protein having *n* residues regardless of its amino acid sequence.

Although an infinite number of ϕ ψ dihedral angle combinations are theoretically possible when building a peptide backbone, a finite set is typically used for computational expediency. In preferred embodiments, the finite set of allowed dihedral angles represent the most energetically favorable and most populated regions of the Ramachandran plot. An illustrative example of a finite set is the six ϕ ψ dihedral angles that are used in the most preferred embodiments: (-65, -42); (-123, 139); (-70, 138); (-87, -47); (77, 22); and (107, -174). Regardless of the set of ϕ ψ dihedral angles, bond lengths and bond angles are fixed to standard values and the peptide torsion angle ω is fixed to 180° for all residues.

If an α-carbon only representation is used, a finite set of dihedral angles (as defined by $C\alpha_i$-$C\alpha_{i+1}$-$C\alpha_{i+2}$-$C\alpha_{i+3}$) may be similarly calculated by generating a Ramachandran-like plot for Cα dihedrals that are found in solved protein structures. In this version of the buildup procedure, the Cα-Cα distance is fixed at preferably about 3.8 Å.

As described previously, the initial set of ensemble structures represents an exhaustive enumeration of the possible tertiary folds for a *n* residue protein. To this end, any suitable sampling technique may be used. Because of the general preference for reduced representation, methods developed for general polymers may also be used with the inventive method with little or no modification. Several such techniques are described in greater detail in the Experimental Section. However, because of its superior sampling efficiency, the method referred to as the Continuous Configuration Boltzmann Biased Direct Monte Carlo Method as described by Sadanobu and Goddard in J. Chem. Phys. **106**: 6722 (1997) and incorporated herein by reference in its entirety, is most preferred.

For the purposes of clarity, the residue buildup procedure will be explained using a peptide backbone representation from this point forward. However, it is to be understood, that any representation may be used and that persons skilled in the art would readily be able to adapt the described methods for the particular representation.

The first step in the residue by residue build up procedure involves selecting a residue position from the amino acid sequence of the protein. The choice of the initial residue is not critical and may be at any point along the sequence. In preferred embodiments, a residue in the middle of the sequence is chosen. The coordinates for the first residue may be fixed at any point in the available coordinate space using standard bond lengths and angles for N, Cα, C, and O and setting the peptide torsion angle, ω, at 180°.

Once the first residue is chosen, residues are added one at a time until the entire protein is constructed. With the exception of the first residue, ϕ ψ angles must be assigned to each residue that is added to the growing fragment. Although any suitable method may be used, a Metropolis based method is generally preferred where the probability of selecting one of the one of the available pairs of dihedral angles is governed by the ratio of the Boltzmann energy for the individual dihedral pair over the sum of the Boltzmann energies for all of the available dihedral pairs. Assuming that the set of dihedral angles is limited to six as in the most preferred embodiments, then the probability is governed by the equation:

$$P_j = \frac{e^{\frac{-E_j}{RT}}}{\sum_{i}^{6} e^{\frac{-E_i}{RT}}}$$

Once the ϕ ψ dihedral angles are determined, the residue is added to the existing fragment. The energy of the fragment with the added residue is then calculated. Although any suitable method for evaluating the energy of the growing fragment may be used, the use of pair wise interaction energies is generally preferred. Moreover, because the non-bonded interactions falls dramatically within a short distance, the use of a specified cut-off distance is also preferred. A suitable cut-off distance from the added residue is between about 8 Å and about 10 Å.

Examples of methods for calculating nonbonded interaction energies include but are not limited to Lennard-Jones 12-6 potentials, Lennard-Jones 8-4 potentials, Morse Potentials, and Exponential-6 potentials. However, the Lennard-Jones 12-6 potential is generally preferred.

Because the energy calculation is to prevent the occurrence of overlapping residues, for computational expediency, it is most preferred to perform the energy calculations based on α-carbon positions only even if a more complete representation of the protein is used to build the model (*i.e.*, peptide backbone). Consequently, in the most preferred embodiments, the nonbonded energy for adding residue *i*, E_i, is given by the sum of its pair-wise interaction energies of the α-carbons of the peptide fragment using the Lennard-Jones 12-6 potential:

$$E_{nb}(R) = E_0\left[\left(\frac{R}{R_0}\right)^{12} - 2\left(\frac{R}{R_0}\right)^{6}\right],$$

wherein R is the distance between the α-carbons of each residue; *i* and *j* are non-adjacent neighbors in sequence; and E_0 and R_0 are set to predetermined values.

If the added residue results in overlapping peptide backbones (as determined by the unfavorable energy of the fragment), protein conformations containing the resulting fragment are no longer pursued. This residue adding procedure in then repeated until the entire protein is constructed.

In preferred embodiments, R_0 is set between about 5 and 6 angstroms for all residue types, and E_0 is set between about 0.1 and 0.2 kcal/mol. In the most preferred embodiment, R_0 is set at about 5.5 angstroms for all residue types; E_0 is about 0.15 kcal/mol. However, the exact values of R_0 and E_0 are not critical to the method and may be set to other values. Moreover, sequence dependency may also be introduced into the energy calculation, if desired. For example, a different E_0 may be used for each of the twenty amino acid residues. In another variation, residues are designated as either hydrophobic and hydrophilic based on known methods and three different energies (E_0's)

are assigned. As with the sequence independent E_0, the precise energy values for the sequence dependent E_0's are not critical and may be determined by any number of heuristic methods known in the art. However, one example of suitable energies are: E_0 = 0.15 kcal/mol for an interaction between two hydrophilic (or polar) residues; E_0 = 0.20 kcal/mol for an interaction between a hydrophilic and a hydrophobic (non-polar) residue; and E_0 = 0.25 kcal/mol for an interaction between two hydrophobic residues.

Although it is not necessary to the practice of the invention, the use of an enrichment technique is preferred. In general, after a residue is added to the fragment, then *m* copies of the fragment is governed by:

$$\text{int}[(z_i/\langle z_i\rangle)/(z_{i-1}/\langle z_{i-1}\rangle)]$$

wherein $z_i = \exp(-E_i/kt)$ which corresponds to the Boltzmann factor for the particular fragment in which the *i*th residue has just been added (thus resulting in an *i* residue fragment wherein i > 1); $\langle z_i\rangle$ is the accumulated average Boltzmann factor for all the fragments in the ensemble having *i* number of residues; z_{i-1} is the Boltzmann factor for the particular fragment without the *i*th residue; and $\langle z_{i-1}\rangle$ is the accumulated average Boltzmann factor for all fragments in the ensemble having *i-1* number of residues. Each copy is then grown independently.

The enrichment method incorporates a novel memory saving algorithm which is further described in the Experimental Methods section. Briefly, after a complete protein chain is constructed, a chain generation counter, $F_i = m_i$ is calculated for each residue *i*. This protein chain is stored in memory. Starting at the end of the protein chain, the method backtracks through the residue addition steps in the opposite order in which the residues were added until a value of $m_k > 1$ is found for some residue *k*. The portion of the protein from residue *i*=1 to *i*=*k* is replicated and a new offspring protein

chain is constructed by adding residues to the replicated fragment. When the new offspring protein chain is completed, enrichment factors for each residue in the offspring are calculated and the chain generation counter at residue *k* in the parent chain is reduced by 1. The method again backtracks through the residue addition steps in the parent chain until another residue whose enrichment factor is greater than one is found. This procedure continues until the chain generation counter is 0 or 1 for each residue in the parent.

Ensemble Reduction

Once the ensemble of initial candidate tertiary structures are generated for a *n* residue protein, the next step in the process is to reduce the number of candidate structures. The first class of recognition filters is sequence independent and generally relate to the observation that proteins tend to be compact structures. This characteristic may be quantified using a number of known methods including the radius of gyration and moments of inertia. However, because the values are more easily calculated, methods based on using the radius of gyration are generally preferred.

A powerful sequence independent filter is to exclude those structures that do not have native-like radius of gyration values. In preferred embodiments, only those members of the ensemble with a radius of gyration between $\delta_1 \cdot R_{min}$ and $\delta_2 \cdot R_{min}$ are selected wherein R_{min} is determined by the following

$$R_{\min}(n_r) = -1.26 + 2.79 n_r^{1/3},$$

n_r is the number of amino acid residues in the protein, and δ_1 and δ_2 are predetermined values. Preferred values for δ_1 are between about 0.9 and 1 with about .95 being the most preferred. Similarly, preferred values for δ_2 are between about 1.4 and about 1.5 with about 1.3 being the most preferred.

However, the exact values of δ_1 and δ_2 are not critical and their values may be outside of this range.

This relatively simple calculation is able to eliminate approximately 70% of the structures in the initial ensemble set when δ_1 and δ_2 are set to preferred values. Because native protein structures have radius of gyration values which are at least about 10 to 15% above this minimum threshold, the likelihood of the correct candidate structure being eliminated by this criterion is minimal.

A second class of recognition filters is sequence dependent which are either based on distance constraints or heuristic observations of native protein structures. Illustrative examples of distance constraints include but are not limited to the spatial proximity necessary between non-adjacent residues in sequence to satisfy disulfide bond requirements, metal coordination site requirements, and NMR derived NOE constraints. When known distance constraints are applied, only those members that satisfy the relative spatial arrangement of two or more residues are selected for further consideration.

Although these distance-based constraints are used in the ensemble reduction phase in preferred embodiments, they may also be used as part of the ensemble generation process. For example, if at least one of the necessary distance constraints are not met in the growing fragment, such as two cysteine residues which form a known disulfide bridge not being spatially adjacent (*i.e.*, Cα's within about 9 Å), then that cluster based upon the particular fragment would be terminated. However, it should be apparent that the end result is the same regardless of whether the distance based constraints are applied during the initial ensemble generation or whether the distance based constraints are applied after all possible tertiary folds for the *n* residue protein are generated.

An example of a heuristic based criteria is the observation that hydrophobic residues in the protein must be able to access the protein core, and charged residues in the protein must be able to access the protein surface. Because of the extensive calculations involved and the necessity of having more refined candidate structures that includes at least pseudo-sidechain positions, the tendency for amino acid residues to partition in this manner has not been generally adapted for routine use in *ab initio* modeling protocols.

To remedy these drawbacks, the present invention describes two procedures based only on Cα positions. The implementation of each procedure requires designating amino acid residues into at least two categories, hydrophobic and hydrophilic (or non-hydrophobic). Any standard method for assigning hydrophobicities may be used including but are not limited to methods described by Kyte & Doolittle, Kauzmann, Nozaki & Tanford, Eisenberg, Chothia, and Huang & Levitt. In preferred embodiments, hydrophobic and hydrophilic residues are defined as described by Huang *et al.* in J. Mol. Biol. **252**: 709-720 (1995) and J. Mol. Biol. **257**: 716-722 (1996), both of which are incorporated by reference it their entirety herein.

The first method measures the ability of a hydrophobic residue to access the core. In a preferred embodiment, the center of mass of the candidate protein structure is calculated from the Cα positions. If a residue is hydrophobic, then a vector from the center of mass to the residue is constructed. A hydrophobic residue is deemed incapable of accessing the protein center and thus receives an energy penalty if another residue within a cutoff value is present between the center of mass and the hydrophobic residue. Preferred values for the cutoff is between about 0.4 and 0.6 Å, and more preferably about 0.5 Å. Generally this penalty is expressed energetically by assigning a positive energy value, preferably about 4 kcal/mol, for any hydrophobic residue between 0.95 R_{min} and 1.3 R_{min} of the center of mass. R_{min} is a

function of the number of amino acid residues in the protein, n_r, and is defined as

$$R_{\min}(n_r) = -1.26 + 2.79 n_r^{1/3},$$

In more preferred embodiments, in addition to a hydrophobic penalty, hydrophobic residues that are able to access the core are assessed a favorable energy value. Illustrative examples of such a scheme include but are not limited to:

i) assigning a positive energy value, preferably about 4 kcal/mol, for any hydrophobic residue between 0.95 R_{min} and 1.3 R_{min} of the center of mass which is unable to access the center;

ii) assigning a negative energy value, preferably about -15 kcal/mol, for any hydrophobic residue within 0.95 R_{min} of the center of mass; and,

iii) assigning a less negative energy value than in ii), preferably about -10 kcal/mol, for any hydrophobic residue between 0.95 R_{min} and 1.3 R_{min} of the center of mass which is able to access the center.

In an even more preferred embodiment, the ability of hydrophilic residues to access the surface is evaluated in addition to the ability of the hydrophobic residues to access the core. The hydrophobic residues are treated as described above. As for the hydrophilic residue, instead of a vector, a cone is constructed from the center of mass as its vertex, preferably with a cone angle of 5 degrees, that encompasses the hydrophilic residue. A hydrophilic residue is deemed incapable of accessing the surface if the extension of the cone beyond the hydrophilic residue contains another residue. An illustrative set of energy parameters are:

i) assigning a positive energy value, preferably about 4 kcal/mol, for any hydrophobic residue between 0.95 R_{min} and 1.3 R_{min} of the center of mass that is unable to access the center;

ii) assigning a negative energy value, preferably about -15 kcal/mol, for any hydrophobic residue within 0.95 R_{min} of the center of mass;

iii) assigning a less negative energy value than in ii), preferably about -10 kcal/mol, for any hydrophobic residue between 0.95 R_{min} and 1.3 R_{min} of the center of mass that is able to access the center; and,

iv) assigning a positive energy value, preferably about 2 kcal/mol, for any hydrophilic residue that is incapable of accessing the surface.

In another variation of the first method, the center of mass of the protein and R_{min} are calculated as described above but uses a simplified scoring system based upon the identities of particular residues. If a particular hydrophobic residue is within a predetermined distance from the center of mass, then it receives a score of -1. However, if the residue is outside the predetermined distance, then it receives a penalty score of +2. The hydrophobic distance cutoffs are: 1.2 R_{min} for phenylalanine and isoleucine; 1.25 R_{min} for leucine and valine; and 1.3 R_{min} for cysteine. In contrast, if a particular hydrophilic residue is within a predetermine distance from the center of mass, then it receives a penalty score of +2. However, if the hydrophilic residue is outside of this distance, then it receives a score of -1. The hydrophilic distance cutoffs are: 0.85 for aspartic acid; 0.8 for asparagine, glutamine, glutamic acid, lysine, proline, and serine; and 0.75 for arginine. If desired, an even more elaborate method may be used wherein the environment of the nearest sequence neighbors are taken into account or wherein a smooth sigmoid function replaces the strict distance cutoffs.

In each of the above described methods, the energy of the candidate conformation is calculated by summing the energies of the individual residues. A predetermined number of conformations having the lowest energies is then selected for further refinement.

Developed by Huang *et al.* at Stanford, the second method attempts to measure a hydrophobic fitness score by counting the number of hydrophobic contacts ("hydrophobic term") and the degree to which hydrophobic residues are buried ("burial term"). The procedure works with either an all Cα or with a Cα and pseudo Cβ protein representation. Although the method will be described in terms of the latter, it may be readily adapted to work with an a Cα representation by substituting Cα-Cα distances instead of pseudo sidechain distances.

The two terms, Hydrophobic Term and the Burial Term, are defined as follows:

$$HydrophobicTerm = \frac{\sum_i (H_i - H_i^{chance})}{n}$$

$$BurialTerm = \frac{\sum_i B_i}{n} .$$

where for each hydrophobic sidechain *i* (as defined by Cβ), H_i is the number of neighboring hydrophobic sidechains within a specified distance from *i*, preferably about 7.3 Å, H_i^{chance} is the number of hydrophobic sidechain contacts which would be expected to occur strictly by chance, and B_i is the total number of neighboring sidechains within a specified distance from residue *i*, preferably about 10 Å. The hydrophobic fitness score, HF, is defined as,

HF = (Hydrophobic Term) x (Burial Term).

If desired, candidate conformations may be minimized as a function of HF, while preserving the overall tertiary conformation. In this procedure, each of the hydrophobic sidechains is directed towards the center of mass of the

protein while directing the hydrophilic residues away from the protein center. The center of mass of the hydrophobic residues is then calculated. Sidechains of a specified length, preferably about 3 Å, are then placed on each hydrophobic residue and all of these sidechains are directed towards the hydrophobic center of mass. Hydrophobic sidechains not within 6 Å of the hydrophobic center are directed away form the protein center of mass and the hydrophobic center of mass is recalculated. This step ensures that only those residues in a single, compact hydrophobic core are included. Sidechains for hydrophilic amino acids are directed away from the hydrophobic center of mass. At the end of the run, the candidate structures are ordered according to their HF score and a predetermined number or percentage of members of the ensemble having the best HF scores are then selected.

Ensemble Refinement

The last major step in the inventive *ab initio* modeling procedure is further refining the remaining candidate structures. In general, approximately between about 100 and about 1000 candidate structures are expected to remain in the ensemble at this point.

High level refinement may be carried out using any of the known molecular mechanics minimization and molecular dynamics simulation methods. Full atom sidechains may be added to the backbone template structures in a computationally efficient manner using sidechain rotamer libraries. Illustrative examples of suitable rotamer libraries include but are not limited to those described by Ponder and Richards, J. Mol. Biol. **193**: 775-791 (1987) and Dunbrack and Karplus, J. Mol. Biol. **230**: 543-574 (1993). In preferred embodiments, the energy of solvation (the interaction of the structure with the solvent molecules in solution) is considered either explicitly or through known statistical mechanical formulations. Because the numbers will be sufficiently small at this stage, simulations may be carried out on all

remaining members of the ensemble to determine the minimum energy configuration.

EXAMPLE 1

Potential Solution to Levinthal's Paradox

The size of the complete set of topologically distinct conformations for a *n* residue polypeptide was determined by generating ensembles of protein structures using the Generic Protein Direct Monte Carlo method. This method applies the CCBB direct Monte Carlo growth technique in conjunction with a generic energy function and peptide representation that treat all amino acid types identically. Because the energy expression is not dependent on amino acid sequence identify, a generic protein ("GP") ensemble contains a highly diverse set of self-avoiding protein conformations.

In order to determine how large an ensemble must be to sample all self-avoiding folded topologies for a *n* residue protein, test sets were compiled of about 20 native proteins and protein domains for residue number, n = 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 80, and 100 residues. The size of the complete set of topologically distinct self-avoiding protein folds was estimated by determining how many GP conformations were required to find a near-native conformation for all each member of the test set (see also Figure 3). The result of this experiment is that the number required in the GP ensemble scales as approximately 1.2^n which is substantially slower than the benchmark figure of 3^n. Figure 4 is a graphical illustration of this point.

The Levinthal Paradox is founded on the assumption that there are 3^n conformation states for a *n* residue polypeptide. Because it was generally assumed that a polypeptide could not sample more than 10^{13} conformations in one second, sampling all 3^{100} or 10^{48} states estimated for a 100 residue peptide was not believed possible. The paradox arises because despite the

staggering number of potential conformations, proteins nevertheless are able to find the global energy structure within millisecond timescales.

Although a directed reaction pathway was suggested by Levinthal as a potential solution, recent experiments have shown that a restrictive single pathway model is not required. The key reduction in states might also be achieved by a multiple pathway mechanism, akin to a funnel. However, since even the highly ordered native state is only marginally more stable than an unfolded state, these models do not suggest a plausible explanation for how an early folding mechanism precludes the protein from sampling vast regions of unproductive conformation space.

The GP folding studies suggest an alternative solution to the Paradox. For example, it is estimated that there are $3x10^7$ topologically distinct conformations for a 100 residue peptide. Assuming an average sampling rate of one new state per 0.3 nanosecond, it is estimated that $3x10^7$ states (the number of all folded conformations estimated for a 100 residue protein) can be sampled in approximately 10 milliseconds. Thus, proteins up to 160 residues can fold by random sampling in one second. Since domains are typically no larger than 200 residues, it is possible for even multi-domain proteins to fold in sub-second time scales if the nucleation of locally static secondary structure within each domain reduces the net conformational freedom to the equivalent of a 160 residue unconstrained peptide.

Arguments against such a random picture of folding are sometimes based on the experimental observation that fragments excised from proteins and kinetic folding intermediates possess native-like secondary structure. However, such biases are not required for small proteins or domains to fold. For example, 86 amino acid reduced HIV-1 Tat (trans-activator) protein folds to a structure with a well-defined core, yet possesses no secondary structure or disulfide bonds.

The GP folding studies are believed to be consistent with recent experimental findings that protein folding is not generally confined to a single reaction pathway with readily identifiable intermediates. These results provide a plausible explanation of not only multiple folding pathways but also how even large proteins are able to fold to their unique three dimensional conformations in sub-second time scales.

EXAMPLE 2

Restricting the available conformational space to self-avoiding folded topologies provides the crucial reduction in the number of potential conformations in the *ab initio* folding problem. Figure 7 illustrates four proteins in which the inventive method of used: (a) 65-residue segment from the NMR determined structure of the proteolytic fragment from Bacteriorhodopsin (1bct); (b) 65 residue Porcine C5a (1c5a); (c) 80 residue fragment from acyl-coenzyme A binding protein (1aca); and, (d) 80 residue segment from domain four of the N-terminal domain of 70 kD heatshock cognate protein (1hpm04).

The superimposition of the native structure with the corresponding GP equivalent results in the following CRMS values: (a) 65-residue segment from the NMR determined structure of the proteolytic fragment from Bacteriorhodopsin (1bct): the GP structure has a CRMS fit of 5.78 Å ; (b) 65 residue Porcine C5a (1c5a): the GP structure has a CRMS fit of 5.40 Å; (c) 80 residue fragment from acyl-coenzyme A binding protein (1aca): the GP structure has a CRMS fit of 6.12 Å; and, (d) 80 residue segment from domain four of the N-terminal domain of 70 kD heatshock cognate protein (1hpm04): the GP structure has a CRMS fit of 6.14 Å.

Refining the corresponding GP structure results in a structure with the following CRMS values when compared with the native structure: (a) 65-residue segment from the NMR determined structure of the proteolytic

fragment from Bacteriorhodopsin (1bct): refined structure has a CRMS of 4.35 Å; (b) 65 residue Porcine C5a (1c5a): the refined structure has a CRMS of 3.91 Å; (c) 80 residue fragment from acyl-coenzyme A binding protein (1aca): the refined structure has a CRMS of 4.97 Å; and, (d) 80 residue segment from domain four of the N-terminal domain of 70 kD heatshock cognate protein (1hpm04): the refined structure has a CRMS of 4.22 Å.

As illustrated by Figure 7, a GP equivalent to the native folded topology is found using the inventive methods. However, it is evident that native local backbone configurations, such as secondary structure, are not well described. This is due to the crude energy function which does not take into account hydrogen bonding or other sequence specific interactions during the initial stages of the protocol. However, the lack of secondary structure during the ensemble generation process is not disconcerting since methods for including sequence specific information may be incorporated during the various stages of refinement. As also illustrated by the superimposition of the native with the refined structure in Figure 7, local refinement of GP folded topologies can result in conformations that are both globally and locally similar to native protein structures.

As a result, coupled with various selection techniques, the use of the GP method significantly reduces the complexity of the *ab initio* folding problem to one of selecting and/or refining a set of structures by incorporating sequence specific information.

EXPERIMENTAL METHODS

Compilation of native protein sets

In order to increase the number of experimental structures for each peptide length n, we included longer structures truncated at the carbonyl terminus. For example, the test set for $n = 45$ consist of residues 1-45 from available protein structures length 45-49. In instances where the coordinate file

contained more than one set of coordinates for a given structure, the first set was used.

The proteins are identified by either the protein identifiers used by the Brookhaven Databank or the CATH database.

The 20, 25, and 30 residue proteins were constructed using the 20 parent proteins: 1aph, 1cbh, 1chl, 1cld, 1cta, 1dec, 1dfn, 1dmc, 1erp, 1fct, 1ktx, 1pnh, 1ppt, 1sis, 1sxm, 2achB, 2mhu, 2pgd03, 4cpaI, 7znf.

35mers (20): 1aml, 1apo, 1cbh, 1chl, 1dec, 1erd, 1erp, 1ica, 1ktx, 1ltsC, 1olgA, 1pcp02, 1pptA4, 1r0904, 1sis, 1sxm, 2achB, 2pgd03, 4cpaI, 9wgaA1.

40mers (21): 125d, 1aml, 1apo, 1bds, 1eptA, 1erd, 1fc2C, 1hev, 1htrP, 1hymA, 1ica, 1ltsC, 1olg, 1pcp02, 1poxA4, 1r0904, 1res, 1zaq, 2erl, 2pdd, 9wgaA1.

45mers (18): 1ahl, 1atx, 1bia03, 1crn, 1ehs, 1gln03, 1gps, 1hnr, 1huc, 1ilk02, 1iva, 1loeB, 1mylA, 1oma, 1pdc, 1shl, 1yrnA, 2ech.

50mers (21): 1afp, 1bal, 1brbI, 1egf, 1enh, 1fdx, 1hcgB, 1lccA, 1mbe, 1ncfA2, 1ptq, 1raaB2, 1sgp, 1tfi, 1tih, 1tpm, 2atcB2, 2tgf, 3monB, 4sgbl, 6insE.

55mers (23): 1aaf, 1amg02, 1amy02, 1bbo, 1bpb01, 1ctm02, 1d66A, 1drs, 1fca, 1gfc, 1hcc, 1lyaB1, 1pdnC2, 1pgb, 1pnrA1, 1prlC, 1ysaC, 2baa02, 2mev4, 2reb02, 3aahB, 3ovo, 5pti.

60mers (23): 1ata, 1cseI, 1dem, 1fxrA, 1gatA, 1hdp, 1hfh, 1igd, 1isuA, 1mdyB, 1nra, 1ntx, 1pce, 1pi2, 1r69, 1rhpA, 1rpo, 1scmA, 1sso, 1trlA, 2drpA, 2hntE, 4mt2.

65mers (24): 1ahdP, 1bct, 1bfmA, 1bhb, 1c5a, 1chc, 1cis, 1copD, 1ctf, 1hre, 1hrt, 1kbaA, 1kst, 1mjc, 1mntA, 1napA, 1ocp, 1pse, 1rtnA, 1sap, 1stu, 1wapA, 2cro, 2sn3.

70mers (21): 1bbi, 1bod, 1bpb03, 1cksA, 1ftz, 1fvl, 1gbrA, 1hcqA, 1hma, 1hoe, 1hpi, 1hstA, 1lea, 1neq, 1ntn, 1octC, 1osa02, 1pkp01, 1spbP, 1utg, 2hsp.

80mers (24): 1aba, 1aca, 1apa02, 1bgh, 1ctl, 1cyg03, 1cyi, 1cyo, 1eptB, 1gtrA3, 1hip, 1hpm02, 1hpm04, 1hra, 1lab, 1pba, 1pht, 1poh, 1pyaA, 1tig, 1tiv, 2dln01, 2fxb, 2gcr01.

100mers (22): 1aj, 1ab2, 1acx, 1bet, 1cmbA, 1etc, 1fd2, 1fkb, 1fus, 1hks, 1hrc, 1ltsD, 1onc, 1pal, 1put, 1thx, 1tlk, 1ycc, 2atcB, 2cdv, 2imn, 2pna.

Generic Protein Structure Generation

Using the Generic Protein Direct Monte Carlo method, a complete set of peptide backbone coordinates for a protein is constructed by consecutively adding residues in one of six possible conformation states to a single residue which is selected at random. In preferred embodiments, the first residue is at the center of the peptide sequence. The six possible conformations correspond to six pairs of ϕ ψ dihedral angles are chosen because they represent the most energetically favorable and most populated regions of the Ramachandran plot. However, any number energetically favorable ϕ ψ pairs may be used.

An illustrative example of six pairs of ϕ ψ dihedral angles are: (-65, -42); (-123, 139); (-70, 138); (-87, -47); (77, 22); and (107, -174). Bond lengths and angles are fixed to standard values and the peptide torsion angle ω is fixed to 180° for all residues.

During buildup procedures, the probability of selecting one of the available pairs of dihedral angles is governed by:

$$P_j = \frac{e^{-E_j/RT}}{\sum_i e^{-E_i/RT}} .$$

The addition energy, E_i, of a single residue is given by the summation of its pair-wise interaction energies with each residue in the peptide fragment. For all residues types, the energy of a residue pair is:

$$E_{nb}(R) = E_0\left[\left(\frac{R}{r_0}\right)^{12} - 2\left(\frac{R}{R_0}\right)^{6}\right],$$

wherein R_0 is set at 5.5 angstroms for all residue types; E_0 is 0.15 kcal/mol; R is the distance between the α-carbon of each residue; and *i* and *j* are not adjacent neighbors in sequence.

In preferred embodiments, energetically favorable addition steps are replicated by a factor m which is equal to

$$\text{int}[(z_i/\langle z_i\rangle)/(z_{i-1}/\langle z_{i-1}\rangle)]$$

wherein $z_i = \exp(-E_i/kt)$. Once a complete chain has been constructed, z_i values are calculated for each residue *i* in the completed chain. Initial values of $\langle z_i\rangle$ are based on values derived from 50 pre-completed polypeptide chains. To avoid replication factors that lead to the same basic fold, the value of $(z/\langle z\rangle)$ is set to 1 for the fixed central residue and at the N-terminal and C-terminal residues.

Once the enrichment factors are calculated for the just-completed chain, replication begins. At each residue *i* in the new polypeptide, if $m_i > 1$, then the fragment up through the growth of residue *i* is replicated ($m_i - 1$) times. For example, if $m_i = 2$, then one replication is performed since the original chain counts as one copy. Each of these replicated fragments is extended into complete polypeptide as described previously. These completed polypeptides are subject to replication at any residue in the freshly added growth where $m_i > 1$ for the respective new polypeptide chains. In this manner, the growth of a single polypeptide chain leads to multiple generations of polypeptide chains that contain a central fragment from the parent polypeptide.

In most preferred embodiments, a novel memory saving algorithm is used during the enrichment/replication stage. Once a complete polypeptide is constructed, values of m_i are determined. The residue addition steps are backtracked in the opposite order in which the residues were added until a residue *k* is found for which $m_k > 1$. The protein fragment which incorporates all of the addition steps through the addition of residue *k* is replicated and an offspring polypeptide is created by adding to the newly-replicated fragment. Enrichment factors are calculated for the offspring chain and the value of m_k for the parent chain is reduced by one since one of the replications that was to take place at this residue has been completed.

The backtracking through the parent chain continues until the value of m for each of the residues in the parent chain is 0 or 1. In this manner, each of the original values of m_i in the parent and every subsequent offspring is replicated ($m_i - 1$) times, while the coordinates of a single polypeptide is stored in memory. When this occurs, a new parent is constructed as previously described.

In this manner, biased only van der Waals packing energy, a diverse ensemble of non-overlapping peptide chains with realistic peptide backbone

geometries can be rapidly generated. For example, more than 10^6 conformations for a 50 residue protein was generated in one CPU day on a single processor Silicon Graphics Inc. R10000 workstation.

Native Structure Search and Topology Verification

The GP generated structure and the native structure were optimally superimposed and the root mean squared deviation of the α-carbons ("CRMS") between the structures was calculated. Topological equivalence was determined by a rigorous, automated minimization procedure. Each α carbon in the candidate GP structure was tethered to the coordinates of the corresponding α carbon in the optimally superimposed native structure with a 5 kcal/mol harmonic constraint energy. Minimization was performed on the constrained GP backbone using the Dreiding force field parameters which is described in Mayo *et al.*, J. Phys. Chem. **94**: 8897 (1990) and which is incorporated herein by reference.

If the GP structure and the native structure are topologically equivalent, then during minimization, the GP structure should follow a direct trajectory toward the native structure. In other words, the GP structure should quickly minimize to the native coordinates. Topology differences are easily observed by the inability of the GP structure to minimize to the native coordinates since the force filed parameters do not permit covalent bond breakage in the peptide backbone or allow cooperative movements between non-local residues. The minimization trajectories demonstrate the conformational dynamics for a GP structure to assume a native fold.

DISTANCE CONSTRAINED METHODS

One implementation of the GP generation method which incorporates distance constraints is as follows. This modified method is used whenever an approximate distance between at least two residues is known. Because only structures which comply with this constraint are generated, the

implementation of the distance constraint is a powerful tool in reducing the number of candidate structures.

In addition to protein structure prediction, the distance constrained GP method may be used with automated procedures for NMR peak sequence assignments. For example, using the distance constraints from a few unambiguously assigned peaks, GP conformations may be used to assist in the assignments of other peaks, which in turn can be used to further reduce the number of candidate conformations. This process can be reiterated until all observed peaks are assigned. A similar procedure may be developed for any other experimental or theoretical process which gives pair-wise or other structure information. Illustrative examples include but are not limited to spectroscopic labeling experiments, or x-ray intensity data wherein the Patterson function from Fourier transformation is used directly with the assignment of phases.

It is preferred that a Cα model is used during the initial stages with the Cα-Cα bond length, b, set to 3.8 Å and the Cα-Cα-Cα angle, θ, set to 120°. As described previously, the Cα dihedrals, ϕ, may either be determined using a Cα version of a Ramachandran plot or a crude six state residue representation of ϕ = to 0°, 60°, 120°, 180°, 240°, or 300°.

Because a Cα representation is used, the coordinates for the first three residues are fixed from standard Cα bond length b and Cα-Cα-Cα angle θ. For example, if the coordinates of the first residue are set to 0,0,0, then the coordinates for the first three residue would be as follows:

$C\alpha_i$= (0.00, 0.00, 0.00).

$C\alpha_{i+1}$= (3.80, 0.00, 0.00).

$C\alpha_{i+2}$= (5.70, 3.29, 0.00).

Once the coordinates for the initial three contiguous residues are determined, the remaining coordinates are determined by the Cα dihedral ϕ. As previously described, the probability of selecting one of the available pairs of dihedral angles during build up procedures is governed by:

$$P_j = \frac{e^{-E_j/RT}}{\sum_i e^{-E_i/RT}} .$$

The addition energy, E_i, of a single residue is given by the summation of its pair-wise interaction energies with each residue in the peptide fragment. For all residues types, the energy of a residue pair is:

$$E_{nb}(R) = E_0\left[\left(\frac{R}{r_0}\right)^{12} - 2\left(\frac{R}{R_0}\right)^{6}\right],$$

wherein R_0 is set at 5.5 angstroms for all residue types; E_0 is 0.15 kcal/mol; R is the distance between the α-carbon of each residue; and *i* and *j* are not adjacent neighbors in sequence.

The following is one implementation of the inventive method which takes one or more known distances into account as the structures are being generated. As it will be described in more detail below, the probability of selecting a residue addition step which would result in a structure which cannot meet the distance constraint is zero.

Figure 8 is a representation of a peptide fragment in which residues *i*-1, *i*, *i*+1, *i*+2, *i*+3, and *i*+4 all line in the same plane. In other words, ϕ_{i+1}, ϕ_{i+2}, and ϕ_{i+3} all are either 0° or 180°. For the purposes of illustration, a

cylindrical coordinate system is assumed with the z-axis travelling through the bond between residue *i*-1 and residue *i*, and the z axis-origin at residue *i*-1. The radial axis, r, represents the perpendicular distance to the z-axis from any point in space. Given this coordinate system, it is possible to determine the maximum radial distance, r_{max}, that residue (*i*-1)+n may be from the z-axis for a given value of z. The solid line connecting the possible in-plane coordinates for residue *i*+4 in Figure 8 shows the boundary of r_{max} as a function of z for n = 5.

Given this peptide representation, where $\theta = 120°$ and b = 3.8 Å, it is possible to write a general equation for r_{max} in terms of z and n. When n is even, then for $z \geq 3b/2$, $r_{max} = r_{peak} - (\tan 30°)(z-(3b/2))$ and for $z < 3b/2$, $r_{max} = r_{peak} + (\tan 30°)(z-(3b/2))$. Here $r_{peak} = (n-1)(b\sin 60°)$, and z lies between $\{z_{min}, z_{max}\} = \{(-3b/4)(n-4), (3b/4)(n)\}$. When n is odd. then for $z \geq b$, $r_{max} = r_{peak} - (\tan 30°)(z-b)$ and for $z < b$, rmax = $r_{peak} + (\tan 30°)(z-b)$. Consequently, $r_{peak} = (n-1)(b\sin 60°)$, and z lies between $\{z_{min}, z_{max}\} = \{(-b/4)(2+3(n-5)), (b/4)(4+3(n-1))\}$. By using these relationships, it is possible to determine the furthest position in (z, r) space that residue (*i*-1+n) may be placed from residues *i*-1 and *i*.

For example, consider a case in which a distance between residue *j* and *k* are known. For the purposes of illustration, assume that in the existing fragment, residue *j* has been fixed, residue *k* has yet to be added, and the residue to be added is residue *i*. Using the above equations, as residue *i* is to be placed, the furthest position in (z, r) space residue *k* can be placed from the position of residue *i*-1 and residue *i* may be determined, given the sequence separation $n = k-(i-1)$.

For the purposes of illustration, assume that the distance between residue *j* and residue *k* is such that the maximum allowable that residue *k* may be placed from residue *j* is 6.58 Å which is equivalent to the distance that may

be traversed in two residue addition steps (*i.e* $2b(\sin\theta/2) = 6.58$ Å). In other words, placing residue *k* within 6.58 Å of residue *j* is equivalent to requiring that residue *j* lies in allowed (z, r) space for residue *k*+2. Hence, as the possible locations for residue *i* is examined, residue *j* must reside in allowed (z, r) space for the sequence separation n=*k*+2–(*i*–1) to also satisfy the distance constraint between residues *k* and *j*. If torsion ϕ_i places residue *i* in a location outside of the allowed (z, r) space, then the distance constraint cannot possibly be satisfied, and the probability of selecting this torsion would be zero.

As it can be seen from the above description, it is useful to associate a constraint bond order (o_b), which represents the number of residue addition steps required to span the constraint distance. Distances less than 3.8 Å may be spanned by a single addition step of length b, hence the o_b for this distance is 1. Distances up to $2b(\sin\theta/2) = 6.58$ Å may be spanned by two residue addition steps. In other words, for $3.8Å < d < 6.58$ Å, $o_b=2$. Because most know distance constraints will be either derived from disulfide bond formation or NOESY NMR data, typical residue-residue distance constraints will be between about 3.8 Å and about 7.4 Å. Such distance constraints may be treated with $o_b=2$, until the placement of the final residue in the constrained pair (*i.e.*, placing residue k in the example above), where the actual distance constraint may be used in place of the above equations for z and r with n=3.

Although the above example was illustrated with only one known distance, it is possible that a set of distance constraints affecting several residues in the protein sequence may be known. In the simplest case, a distance constraint is a *first order* constraint. This is the example previously discussed where residues *j* and *k* are constrained by some distance and residue *j* is place prior to residue *k* and residue *i* is between *j* and *k* in sequence such that

$$i-j \geq k-i+o_b-2.$$

A *second order* constraint exists when two distance constraints are coupled such that both distance constraints cannot be satisfied by treating them independently as first order constraints. An example of a second order constraint arises, for example, in the following situation where the distance between residues 5 and 39 are constrained to be within o_b=2 and the distance between residues 17 and 39 are also constrained to be within o_b=2. The distance between residues 5 and 17 is considered a second order constraint with a o_b=4 since it is necessarily dependent on the distance constraints between residues 5 and 39 and the residues 17 and 39. In other words, in considering the possible placement of residue i, both the first and second order constraints must be satisfied. In the above example, for residues $i-5 \geq 17-i+o_b-2$ (*i.e.*, residues 12, 13, 14, 15, 16, and 17), residue 5 must lie in allowed (z, r) space for $n = 17+o_b-(i-1)$, where o_b=4.

In summary, a list of inter-residue distance constraints (along with corresponding bond orders, o_b, is inputted along with N, the total number of residues in the protein. The protein is then constructed residue by residue by moving to the right in sequence, beginning with the initial three N-terminal residues. For each step thereafter, the energy of each possible ϕ angle is evaluated. If a distance constraint cannot be satisfied by a structure including the candidate torsion, then the probability of selecting this torsion angle is set to zero. If the distance constraint may be satisfied, then the probability of selecting this torsion is dependent on the Boltzmann energy as described previously.

At a given residue addition step *i*, if no torsion satisfies the constraint conditions, then polypeptide is re-grown from residue *i*-4, in attempt to satisfy this constraint. The number of times this "backtracking" is performed

may be determined by the user. In a preferred implementation, one "backtrack" is allowed before the entire polypeptide is discarded and a new polypeptide grown from the initial three N-terminal residues.

If desired, a "lookahead" strategy may also be used where the probability of selecting a torsion angle for residue *i* is biased by the placement of residue *i*+1. In this implementation, for a particular torsion candidate for residue *i*, potential torsions for residue *i*+1 are also explored to determine if constraints for residue *i*+1 may be satisfied if the particular torsion for residue *i* is chosen. If this "lookahead" determines that there is no torsion for residue *i*+1 which satisfies the constraints on residue *i*+1, then probability of selecting the particular candidate for residue *i* is set to zero. In preferred embodiments, favorable fragments are not replicated as in non-distance constrained generation methods.

SAMPLING METHODS

The sampling methods that are used are variations that are described by Sadanobu and Goddard, J. Chem. Phys. **106**: 6702 (1997) which is incorporated in its entirety herein. Although the sampling methods described by Sadanobu and Goddard were developed for polymers, they are readily adapted to proteins, especially to calculations where a reduced representation of proteins is used.

The Polymer Model

A united atom model as described by Rychaert and Bellemans, Faraday Discuss. Chem. Soc. **66**: 96 (1977) is used to represent the polymer or peptide backbone. If desired, the equations maybe adapted to include amino acid sidechains. In the model, each atom *i* in the chain is characterized by a Lennard-Jones 12-6 potential as in (1)

$$E_{LJ}(r_{ij}) = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right], \quad |i-j| \geq 4 \tag{1}$$

where ε/k_B = 72K, σ = 0.3923 nm, and r_{ij} is the distance between i^{th} and j^{th} atoms.

In addition, the torsion potential in (2) is included

$$\frac{E_t(\phi_i)}{k_B} = \sum_{n=0}^{5} a_n(\cos\phi_i)^n \tag{2a}$$

where

$$a_0 = 1.157,\ a_1 = 1.515,\ a_2 = -1.636,$$
$$a_3 = -0.382,\ a_4 = 3.271,\ \text{and}\ a_5 = -3.927 \tag{2b}$$

Here ϕ_i is i^{th} torsion angle, and the geometry properties are taken as

$$\textit{bond length}: \ l = 0.153\text{nm} \tag{3}$$
$$\textit{bond angle}: \ \vartheta = 70.53°$$

which corresponds to a carbon-carbon-carbon angle of 109.47°).

The total Hamiltonian has the form

$$H[\{\phi_i\}] = \sum_{i=5}^{N}\sum_{j=1}^{i-4} E_{LJ}[r_{ij}(\{\phi_i\})] + \sum_{i=4}^{N} E_t(\phi_i), \tag{4}$$

the results of which will be quoted in terms of a reduced temperature

$$T_r = \frac{k_B T}{\epsilon}.$$

(5)

The configurational partition function for the model polymer chain consisting of N carbon atoms is defined as

$$Z_N = \int_0^{2\pi} \cdots \int_0^{2\pi} \exp\left[-\beta \cdot H(\{\phi_i\})\right] d\phi_4 \cdots d\phi_N$$

(6)

where $\beta=1/kT$. [Here ϕ_i is the torsion specifying the position of atom i with respect to atoms i-3, i-2, and i-1.]

The Helmholtz free energy A, the potential energy E, and the entropy S, are given by

$$A_N = -\beta \cdot \ln Z_N$$

$$E_N = \int_0^{2\pi} \cdots \int_0^{2\pi} H(\{\phi_i\}) \cdot \exp\left[-\beta \cdot H(\{\phi_i\})\right] d\phi_4 \cdots d\phi_N$$

$$S_N = \frac{(E_N - F_N)}{T}$$

(7)

Simple Sampling (SS)

In the conventional direct Monte Carlo (DMC) method, polymer chains are generated by random step-by-step sampling of torsion angles. A complete N-

mer chain is constructed in sequence where the i^{th} step samples the i^{th} torsion to construct an i-mer chain. Then a new chain is set up and sampled again from scratch. This is referred to as simple sampling (SS). The partition function is evaluated by (8)

$$Z_N = N_c^{-1} \cdot (2\pi)^{N-3} \sum_{1}^{N_c} \exp\left[-\beta \cdot H(\{\phi_i\})\right] \tag{8}$$

Here N_c is total number of chains generated.

The average value $< f >$ of a physical property, $f = f(\{\phi_i\})$, is calculated as in (9),

$$< f > = \frac{\sum_{1}^{N_c} f(\{\phi_i\}) \cdot \exp\left[-\beta \cdot H(\{\phi_i\})\right]}{\sum_{1}^{N_c} \exp\left[-\beta \cdot H(\{\phi_i\})\right]} \tag{9}$$

Independent Rotational Sampling (IRS)

The sampling efficiency of SS-DMC is improved by applying rotationally biased sampling, in which torsions are sampled using a weighting function based on the Boltzmann factor of the torsion energy. This improvement to the simple sampling method is referred to as Independent Rotational Sampling (IRS). In the IRS method, the normalized torsion weighting function (TWF), W_{IRS}, is defined as in (10)

$$W_{IRS}(\phi) = \frac{g_{IRS}(\phi)}{z_{IRS}} \tag{10a}$$

where

$$z_{IRS} = \int_0^{2\pi} g_{IRS}(\phi)d\phi \tag{10b}$$

$$g_{IRS}(\phi) = \exp[-\beta E_t(\phi)] \tag{10c}$$

Torsion angles are generated in accordance with (10a). The partition function for IRS after bias correction is evaluated by (11)

$$Z_N = N_c^{-1} \cdot (z_{IRS})^{N-3} \cdot \sum_1^{N_c} \exp\left[-\beta \cdot \sum_{i-5}^{N} \sum_{j=1}^{i-4} E_{LJ}(r_{ij})\right] \tag{11}$$

W_{IRS} need be calculated only once so that computational work involved in evaluating the partition function involves the Boltzmann factor for the nonbonding energy. With the IRS method, the use of W_{IRS} effectively excludes high torsion energies throughout the MC sampling. Nevertheless, spatial overlaps between non-bonding atoms are inevitable, leading to high

configurational energies. In order to exclude these overlaps, information about the spatial environment in the vicinity of the growing chain end should be introduced into the TWF. The resulting form of the TWF, W^*, is given by (12)

$$W^*(\phi_i; \phi_4,\cdots,\phi_{i-1}) = \frac{g^*(\phi;\phi_4,\cdots,\phi_{i-1})}{z^*(\phi_4,\cdots,\phi_{i-1})} \tag{12a}$$

where

$$z^*(\phi_4,\cdots,\phi_{i-1}) = \int_0^{2\pi} g^*(\phi_i;\phi_4,\cdots,\phi_{i-1})d\phi_i \tag{12b}$$

$$g^*(\phi_i;\phi_4,\cdots,\phi_{i-1}) = g_{IRS}(\phi_i)\cdot\exp\left[-\beta\cdot\sum_{j=1}^{i-4} E_{LJ}(r_{ij})\right] \tag{12c}$$

The form of the partition function after bias correction becomes (13)

$$Z_N = N_c^{-1}\cdot z_{IRS}\cdot\sum_1^{N_c}\left\{\prod_{i=5}^{N} z^*(\phi_4,\cdots,\phi_{i-1}\cdot)\right\} \tag{13}$$

W* must be calculated at every step since it depends on all previous steps. The computation, time for this TWF is approximately proportional to the step

number, i; therefore, this sampling method becomes too expensive for systems containing a large number of atoms.

Continuous Configurational Biased (CCB) Direct Monte Carlo

Another novel sampling method, the Continuous Configurational Biased (CCB) direct Monte Carlo method, is described. In this variation, a cutoff length for non-bonding interactions is introduced into the TWF calculation. On constructing the TWF for the i^{th} torsion, a sphere of radius R_c centered at the $(i - 1)^{th}$ atom position is defined (See Figure 1). The length of R_c should be taken larger than $1 + \sigma$, in order to ensure that all possible atomic overlaps are checked. Boltzmann factors for the non-bonding energy between i^{th} atom and all other atoms inside the cut-off sphere are included in TWF, W_{CCB}, as in (14)

$$W_{CCB}(\phi_i; \phi_4, \cdots, \phi_i) = \frac{g_{CCB}(\phi_i; \phi_4, \cdots, \phi_{i-1})}{z_{CCB}(\phi_4, \cdots, \phi_{i-1})} \tag{14a}$$

where

$$z_{CCB}(\phi_4, \cdots, \phi_{i-1}) = \int_0^{2\pi} g_{CCB}(\phi_i; \phi_4, \cdots, \phi_{i-1}) d\phi_i \tag{14b}$$

$$g_{CCB}(\phi_i; \phi_4, \cdots, \phi_{i-1}) = g_{IRS}(\phi_i) \cdot \exp\left[-\beta \cdot \sum_{j=1}^{i-4} \theta(R_C - r_{ij}) \cdot E_{LJ}(r_{ij})\right] \tag{14c}$$

and $\theta(R)$ is the Heavyside step function

$$\theta(R) = 0 \text{ if } R < 0$$
$$= 1 \text{ if } R \geq 0 \qquad (15)$$

The computation time for W_{CCB} is almost independent of i because the only non-bonding atoms considered are those in the local vicinity of a growing chain end. In addition, the list of atoms inside the cut-off circle for the i^{th} atom is automatically available since all the necessary atomic distances were calculated to obtain the energy at the just previous step. The bias-corrected partition function has the form of (16), which includes the calculation of those non-bonding energies that did not appear in the TWF calculation of (14)

$$Z_N = N_c^{-1} \cdot z_{IRS} \cdot \sum_{1}^{N_c} \left\{ \prod_{i=5}^{N} z_{CCB}(\phi_4, \cdots, \phi_{i-1}) \right\} \cdot$$
$$\exp\left[-\beta \cdot \sum_{i=5}^{N} \sum_{j=1}^{i-4} \theta(r_{ij} - R_C) \cdot E_{LJ}(r_{ij}) \right] \qquad (16)$$

The Continuous Configuration Boltzmann Biased (CC-BB) Method

In the enrichment method for self-avoiding walks on a lattice, once a walk of i - 1 steps is successfully generated by the simple sampling method, this chain continues to be grown up to step i in m_{i-1} different ways. In order to avoid bias, the enrichment factor m_{i-1} is always fixed ahead of the Monte Carlo simulation. The total chain multiplicity, Mi, for step i is defined as

$$M_i = \prod_{j=1}^{i-1} m_j \qquad (17)$$

The chains obtained from a particular first monomer are not statistically independent. Hence, the set of all chains using the same seed as the first monomer are collected together and denoted as a cluster. Each cluster is then given the same weight.

In the replication-deletion procedure (RDP) for a continuous space, chain enrichment is used to achieve a Boltzmann population for the collected chains. Here, M_i is determined at every step as statistically proportional to the ratio of the Boltzmann factor of step (i - 1) to that of step (i - 2), where m_i is not integer. This leads to a high frequency of sampling chains with high energy (caused by nonbonding overlap) which are subsequently deleted in the course of sampling. The partition function is evaluated from the ratio of the total number of generated chains to the number of seeds. To avoid replicating chains too often, a scaling factor p is multiplied by Boltzmann factor. Since the suitable choice of scaling factors is unknown and strongly dependent on chain size and temperature, one determines them in trial and error manner prior to the Monte Carlo simulation. These scaling factors should be fixed ahead of Monte Carlo simulation.

In the Continuous Configurational Biased Direct Monte Carlo (CCB-DMC) method, almost all high energy chains having non bonding overlaps are excluded. As a result, in comparison with replication-deletion procedures, chains are very seldom need to be deleted. However, the sampling distribution is not Boltzmann and low energy chains in the collection can be included with too high a contribution to the partition function. As a result, it

is preferred to extend the chain enrichment to control sampling so that all collected chains make a nearly equal contribution to the partition function.

In this method, the multiplicity, M_i, is determined at every step as proportional to the ratio of the Boltzmann factor of a just-sampled chain to that of the running average value for the chain with same length. The partition function is explicitly calculated as the average of the weighting-bias-corrected Boltzmann factor divided by the chain multiplicity. In the Continuous Configuration Boltzmann Biased (CCBB) method, equation (16) is rewritten in terms of a sum over K clusters as

$$Z_N(K) = K^{-1} \cdot \sum_{C=1}^{K} \zeta_N(C) \tag{18}$$

Denoting $L_n(C)$ as the total number of chains generated for cluster C cluster by using an arbitrary choice for the enrichment factor, the partition function (18) is calculated by (19)

$$\zeta_N(C) = \sum_{n=1}^{L_N(C)} \frac{\zeta_N^n(C)}{M_N^n(C)} \tag{19}$$

$$M_N^n(C) = \prod_{i=1}^{N-1} m_i^n(C) \tag{20}$$

$$L_N(C) = \sum_{n=1}^{L_{N-1}} m_{N-1}^n(C) \tag{21}$$

In CCBB, the chain multiplicity, $M_i^n(C)$, is determined as proportional to the ratio of $\varsigma^n_{i-1}(C)$ to $Z_{i-1}(C\text{-}1)$.

$$Q_i^n(C) = \frac{p \cdot \zeta_{i-1}^n(C)}{Z_{i-1}(C-1)} \tag{22}$$

$$M_i^n(C) = INT[Q_i^n(C)] \quad if\ Q_i^n(C) > 1$$

$$= 1 \qquad\qquad if\ Q_i^n(C) \le 1 \tag{23}$$

This enrichment factor m^n_{i-1} is evaluated from the ratio of M^n_i to M^n_{i-1}.

This procedure always keeps the chain multiplicity approximately proportional to the Boltzmann factor of the chain at the just-previous step.

$$P_i^n(C) = \frac{M_i^n(C)}{M_{i-1}^n(C)}$$

(24a)

$$m_{i-1}^n(C) = INT[P_i^n(C)] \quad if\ P_i^n(C) > 1$$

$$= 1 \qquad\qquad if\ P_i^n(C) \leq 1$$

(24b)

For i < 5 the Boltzmann population of the chain collection is completely satisfied in CCB. Therefore, chain multiplicity is set to unity as

$$M_0^n = M_1^n = \cdots = M_4^n = 1$$

(24c)

The choice of p is arbitrary. Too large a value of p could lead to an exploding number of samples of highly correlated configurations. Too small a value might lead to too few chains per cluster. For the polymer example considered here, p = 1 is used since it results in enriched chains having nearly equal contribution to the partition function.

To obtain an initial guess for the partition function, $Z_i(0)$, a short non-Boltzmann Factor Biased (BFB) run is performed. A Boltzmann Factor Biased (BFB) method is an improved enrichment method, which introduces a configurational-dependent enrichment procedure with correct bias correction and automatic population control. (For this study, 200 chains were sampled prior to BFB sampling.) If the partition function used in the initial guess is too small, an extraordinarily large enrichment factor might occur for clusters after beginning the BFB sampling and ruin the MC sampling. To avoid this, an upper limit can be introduced for the enrichment factor (arbitrarily without any additional bias). For the example reported here, no special controls of

the enrichment factor were needed. Equation (24) gave automatic control of the number of chains generated by BFB sampling. The average number of generated chains per seed atom tended to become large at low temperature. It ranged from 3.5 at the highest temperature to 11.2 at lowest one for N = 400.

CCB Sampling Procedure

Prior to the chain sampling, the torsion energy was calculated for a fixed number of grid points. In this study, 200 equally separated grid points from 0 to 2π)-W_{IRS} were used. The normalized TWF for IRS, is then evaluated using numerical integration for z_{IRS} as in (10b). This uses the auxiliary distribution $P_{IRS}(\phi)$ in (25)

$$P_{IRS}(\phi) = \int_0^{\phi} W_{IRS}(\phi')d\phi' \tag{25}$$

A local Cartesian reference frame is defined for each bond of the chain. The axial trans-formation matrix t_i is

$$t_i = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ \sin\theta\cos\phi_i & -COS\theta\sin\phi_i & \sin\phi_i \\ \sin\theta\sin\phi_i & -\cos\theta\sin\phi_i & -\cos\phi_i \end{bmatrix} \tag{26a}$$

The first atom is set at origin and t_2 and t_3 are set as

$$t_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$t_3 = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ \sin\theta & -\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(26b)

The position vector, R_i, of atom i is calculated as

$$\boldsymbol{R}_i = \boldsymbol{R}_{i-1} + \boldsymbol{T}_i \cdot \boldsymbol{b}$$

$$\boldsymbol{T}_i = \sum_{k=2}^{i} \boldsymbol{t}_k$$

$$\boldsymbol{b}^t = (1,0,0)$$

(26c)

Here, b is the bond vector and T_j is the transformation matrix from the local reference frame on the j^{th} bond to the original reference frame. A random number ξ, uniformly distributed in the interval [0, 1), is drawn and the fourth torsion angle ϕ_4 is obtained by requiring

$$P_{IRS}(\phi_4) = \xi$$

(27)

For i > 4 after sampling the $(i-1)^{th}$ torsion, all non-bond distances are calculated to evaluate the energy and also to define an atom group $\{k\}_i$, whose elements consist of the neighbors of the $(i-1)^{th}$ atom

$$\{k\}_i = \{k \| \boldsymbol{R}_{i-1} - \boldsymbol{R}_k | < R_c;\ 1 \le k < i-5\} \tag{28}$$

The coordinates of all atoms in the list $\{k\}_i$ are transformed into the local reference frame on the $(i - 1)^{th}$ bond by using the inverse matrix $(T_{i-1})^{-1} = T^t_{i-1}$

$$\boldsymbol{R}'_k = (\boldsymbol{T}_{i-1})^{-1} \cdot \boldsymbol{R}_k$$

$$= \boldsymbol{T}^t_{i-1} \cdot \boldsymbol{R}_k \tag{29}$$

In the local reference frame the coordinates of atom i for a trial move at the q^{th} grid point ϕP^q_i is

$$r^t_q = (l\cos\theta,\ l\sin\theta\ \cos\phi^q_i,\ l\sin\theta\ \sin\phi^q_i) \tag{30}$$

which is independent of i if the same type of grid is used for all torsions. $g_{CCB}(\phi_i)$ is evaluated as

$$g_{CCB}(\phi^q_i;\ \phi_4,\cdots,\phi_{i-1}) = g_{IRS}(\phi^q_i) \cdot \exp\left[-\beta \cdot \sum_{\{k\}_i} E_{LJ}(|\boldsymbol{R}'_k - \boldsymbol{r}_q|)\right] \tag{31}$$

Then z_{CCB} and W_{CCB} are evaluated by using the above expression Of g_{CCB}. The auxiliary distribution P_{CCB} is obtained by (32)

$$P_{CCB}(\phi_i) = \int_0^{\phi_i} W_{CCB}(\phi;\phi_4,\cdots,\phi_{i-1})d\phi \tag{32}$$

BFB Sampling Procedure

In the BFB method, the number of chains that will be generated in a cluster cannot be foreseen. Thus, the amount of memory required to store the information for growing branches of the chains in a cluster cannot be predetermined *a priori*. As a result, a memory-saving algorithm is used, in which just one chain is grown at a time.

The complete construction of one chain at time is as follows. For cluster k + 1, the i index starts at i = 4 and increases to i = N. For each such i, each i' from i to N is considered. First, the enrichment factor $m_{i'}$ is determined using (22) and evaluating the running average of the partition function, $Z_{i'}(k + 1)$. For each step i', the chain generation counter $F_{i'}$ is then defined and set to $F_{i'} = m_{i'}$. After calculating F_i from i' = i to i" = N, the calculation then starts at i" = N and work from N back to i. Each $F_{i"}$, is checked to determine if it is greater than unity. When $F_{i"} > 1$ for i" > i - 1, the $i^{"th}$ torsion is sampled once more and $F_{i"}$ is reduced by unity. A new chain is grown from step i" to N and a new value of $m_{i"}$ is evaluated for each step. The same procedure is repeated until there is no $F_{i"}$ larger than unity. At this point the (k + 1)th cluster is completed, and the (k + 2)th cluster can be started. The flow chart of the method is illustrated by Figure 9.

It is to be understood that while the invention has been described above in conjunction with preferred embodiments, the description and examples are

intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims.

What is claimed is:

1. A method for *ab initio* structure prediction for a *n* residue protein backbone, comprising:

selecting a finite set of torsion angles and

generating an ensemble of conformations for the protein backbone wherein the ensemble represents an exhaustive enumeration of self-avoiding backbone conformations for the selected set of torsion angles.

2. The method as in claim 1 wherein the torsion angles are ϕ ψ dihedral angles.

3. The method as in claim 3 wherein the finite set of ϕ ψ dihedral angles include: (-65, -42); (-123, 139); (-70, 138); (-87, -47); (77, 22); and (107, -174).

4. The method as in claim 1 wherein the torsion angles are $C\alpha_i$-$C\alpha_{i+1}$-$C\alpha_{i+3}$-$C\alpha_{i+4}$ dihedral angles.

5. The method as in claim 1 further comprising

calculating the radius of gyration and eliminating those conformations having values outside of the range defined by $\delta_1 \cdot R_{min}$ and $\delta_2 \cdot R_{min}$ wherein R_{min} is

$$R_{\min}(n_r) = -1.26 + 2.79 n_r^{1/3},$$

n_r is the number of amino acid residues in the protein, δ_1 is between about 0.9 and 1.0, and δ_2 is between about 1.4 and 1.5.

6. The method as in claim 1 further comprising

reducing the number of conformations in the ensemble by considering sequence dependent distance constraints selected from the group consisting of disulfide bond requirements, metal coordination site requirements, and NMR derived NOE constraints.

7. A method for determining the three dimensional backbone structure of a protein having an amino acid sequence of R_1-R_2...R_{n-1}-R_n, wherein *n* is the total number of amino acid residues in the sequence, comprising:

selecting a finite set of ϕ ψ dihedral angles and

generating an ensemble of conformations for the protein backbone wherein the ensemble represents an exhaustive enumeration of self-avoiding backbone conformations for the selected set of dihedral angles.

8. The method as in claim 7 wherein the ensemble generation step includes:

(a) selecting a residue *i* within the amino acid sequence;

(b) initiating a growing fragment by determining the three dimensional backbone coordinates for residue *i*;

(c) selecting a residue position that is adjacent to a residue whose three dimensional backbone coordinates have been previously determined;

(d) picking a ϕ ψ dihedral pair from among the finite set using a Metropolis-based sampling method;

(e) determining the three dimensional backbone coordinates for the selected residue;

(f) growing the fragment by adding the selected residue to the fragment;

(g) calculating an addition energy wherein the addition energy is the summation of the pair wise interaction energies of the residues in the fragment;

(h) accepting or rejecting the backbone coordinates for the selected residue based upon an evaluation of the addition energy; and,

(i) repeating steps (c)-(h) until the three dimensional backbone coordinates have been determined for each protein residue.

9. The method as in claim 7 wherein the ensemble generation step includes:

(a) selecting a residue *i* within the amino acid sequence;

(b) initiating a fragment corresponding to residue *i* by determining the three dimensional backbone coordinates for residue *i*;

(c) selecting an existing fragment and selecting a residue position for adding a residue on to the existing fragment;

(d) picking a ϕ ψ dihedral pair from among the finite set using a Metropolis-based sampling method;

(e) determining the three dimensional backbone coordinates for the selected residue;

(f) growing the existing fragment by adding the selected residue thereon;

(g) calculating an addition energy wherein the addition energy is the summation of the pair wise interaction energies of the residues in the fragment;

(h) accepting or rejecting the backbone coordinates for the selected residue based upon an evaluation of the addition energy;

(i) enriching the number of copies of the fragment which includes the coordinates of the selected residue;

(j) repeating steps (c)-(i) until every existing fragment has grown to *n* residues.

10. The method as in claim 9 wherein the addition energy is by

$$E_{nb}(R) = E_0\left[\left(\frac{R}{R_0}\right)^{12} - 2\left(\frac{R}{R_0}\right)^{6}\right],$$

wherein R_0 is between about 5 and 6 angstroms, and E_0 is between about 0.1 kcal/mol and 0.2 kcal/mol for all residue types.

11. The method as in claim 9 wherein the number of copies made in the enrichment step is equal to

$$\text{int}[(z_x/\langle z_x\rangle)/(z_{x-1}/\langle z_{x-1}\rangle)]$$

wherein x is the total number of residues in the fragment which includes the selected residue; $z_x = \exp(-E_x/kt)$ is the Boltzmann factor for the fragment, $\langle z_x\rangle$ is the accumulated average Boltzmann factor for all x length fragments; z_{x-1} is the Boltzmann factor for the fragment without the addition of the selected residue; and $\langle z_{x-1}\rangle$ is the accumulated average Boltzmann factor for all x-1 length fragments.

12. The method as in claim 7 further comprising:

(a) calculating a center of mass for conformation;

(b) determining whether each residue is hydrophobic or non-hydrophobic;

(c) assigning a penalty for each hydrophobic residue found outside a predetermined distance from the center of mass; and,

(d) assigning a penalty for each hydrophilic residue found inside the predetermined distance from the center of mass.

13. The method as in claim 7 further comprising:

(a) calculating a center of mass for conformation;

(b) determining whether each residue is hydrophobic or non-hydrophobic;

(c) assessing a penalty for each hydrophobic residue between about 0.95 R_{min} and 1.3 R_{min} of the center of mass in which another residue is present between the hydrophobic residue and the center of mass, wherein R_{min} is

$$R_{\min}(n_r) = -1.26 + 2.79 n_r^{1/3},$$

and n_r is the number of amino acid residues in the protein; and,

(d) reducing the number of conformations in the ensemble by selecting conformations having the least number of penalties.

14. The method as in claim 7 further comprising:

reducing the number of conformations in the ensemble by calculating an energy for each conformation by:

(a) calculating a center of mass for conformation;

(b) calculating R_{min} using the formula

$$R_{\min}(n_r) = -1.26 + 2.79 n_r^{1/3},$$

wherein n_r is the number of amino acid residues in the protein;

(c) determining whether each residue is hydrophobic or non-hydrophobic;

(d) assessing a first negative energy value for each hydrophobic residue within 0.95 R_{min} of the center of mass;

(c) assessing a second negative energy value for each hydrophobic residue between about 0.95 R_{min} and 1.3 R_{min} of the center of mass in which another residue is not present between the hydrohphobic residue and the center of mass, wherein the first negative value is more negative than the second negative value;

(d) assessing a positive energy value for each hydrophobic residue between about 0.95 R_{min} and 1.3 R_{min} of the center of mass in which another residue is present between the hydrophobic residue and the center of mass; and,

(e) adding the values of the individual residues for selecting the lowest energy conformations.

15. The method as in claim 7 further comprising:

reducing the number of conformations in the ensemble by calculating an energy for each conformation by:

(a) calculating a center of mass for conformation;

(b) calculating R_{min} using the formula

$$R_{\min}(n_r) = -1.26 + 2.79 n_r^{1/3},$$

wherein n_r is the number of amino acid residues in the protein;

(c) determining whether each residue is hydrophobic or non-hydrophobic;

(d) assessing a first negative energy value for each hydrophobic residue within 0.95 R_{min} of the center of mass;

(c) assessing a second negative energy value for each hydrophobic residue between about 0.95 R_{min} and 1.3 R_{min} of the center of mass in which another residue is not present between the hydrohphobic residue and the center of mass, wherein the first negative value is more negative than the second negative value;

(d) assessing a positive energy value for each hydrophobic residue between about 0.95 R_{min} and 1.3 R_{min} of the center of mass in which another residue is present between the hydrohphobic residue and the center of mass;

(e) assessing a negative energy value for each hydrophilic residue that is surface accessible;

(f) adding the values of the individual residues for selecting the lowest energy conformations.

16. A method for determining an alpha carbon backbone structure of a protein having an amino acid sequence of R_1-R_2...R_{n-1}-R_n, wherein *n* is the total number of amino acid residues in the sequence and having at least one distance constraint between two non-adjacent residues in sequence, comprising:

selecting a finite set of Cα dihedral angles and

generating an ensemble of conformations for the protein backbone wherein the ensemble represents an exhaustive enumeration of self-avoiding backbone conformations for the selected set of Cα dihedral angles.

17. The method as in claim 16 wherein the ensemble generation step includes determining the coordinates of the first three N-terminal residues by setting the Cα-Cα bond length equal to 3.8 Å and the Cα-Cα-Cα bond angle equal to 120°.

18. The method as in claim 17 wherein the ensemble generation step further includes:

(a) selecting a residue *i* in the sequence wherein residue *i* is the next amino acid in sequence with respect to the previously placed residue;

(b) selecting a Cα dihedral pair from among the finite set using a Metropolis-based sampling method;

(c) determining the three dimensional backbone coordinates for the selected residue;

(d) determining whether the placing the selected residue would allow the structure to satisfy the distance constraint; and,

(e) rejecting the backbone coordinates if the structure could satisfy the distance constraint or accepting the backbone coordinates if the structure could not satisfy the distance constraint.

19. The method as in claim 18 wherein the backbone coordinates are rejected in step (e) of claim 17 and the ensemble generation step further includes repeating steps (b) through (e) until the backbone coordinates of the selected residue are accepted.

20. The method as in claim 18 wherein the backbone coordinates are accepted in step (e) of claim 17 and the ensemble generation step further includes repeating steps (a) through (e).

21. The method as in claim 16 wherein the generated ensemble is used with a procedure for determining NMR peak sequence assignments.

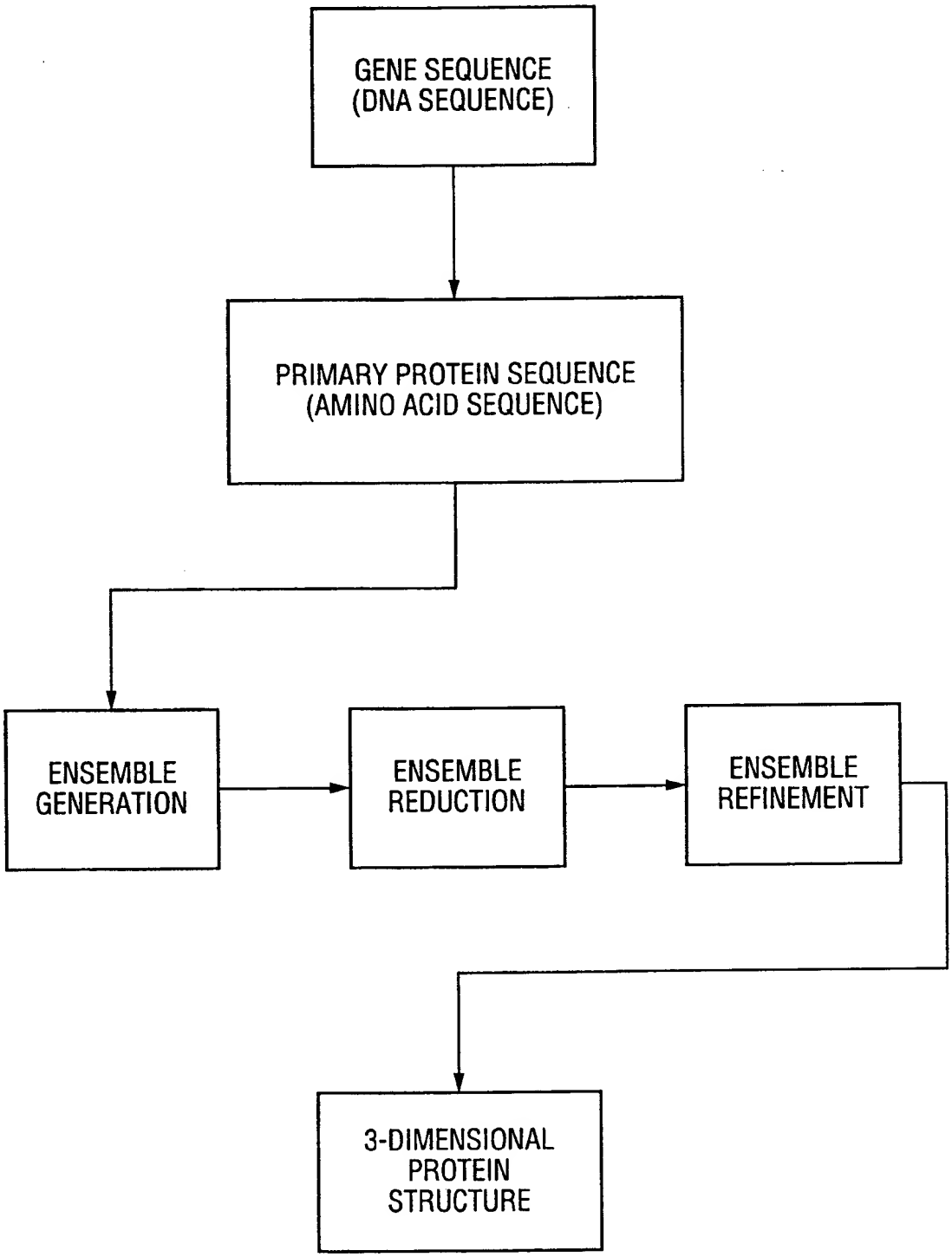

GENE SEQUENCE
(DNA SEQUENCE)
PRIMARY PROTEIN SEQUENCE
(AMINO ACID SEQUENCE)
ENSEMBLE
GENERATION
ENSEMBLE
REDUCTION
ENSEMBLE
REFINEMENT
3-DIMENSIONAL
PROTEIN
STRUCTURE

FIG. 1

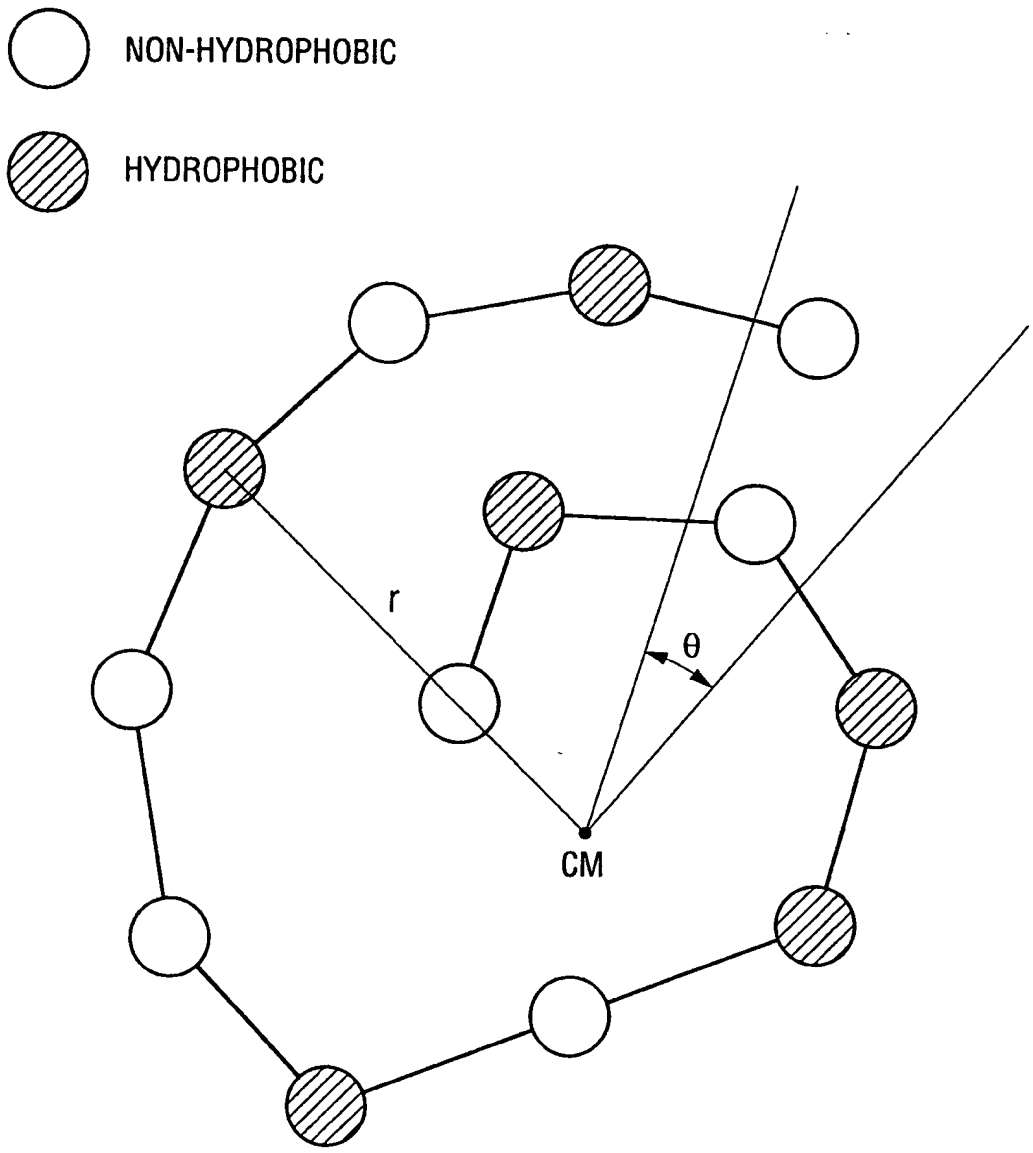


FIG. 2

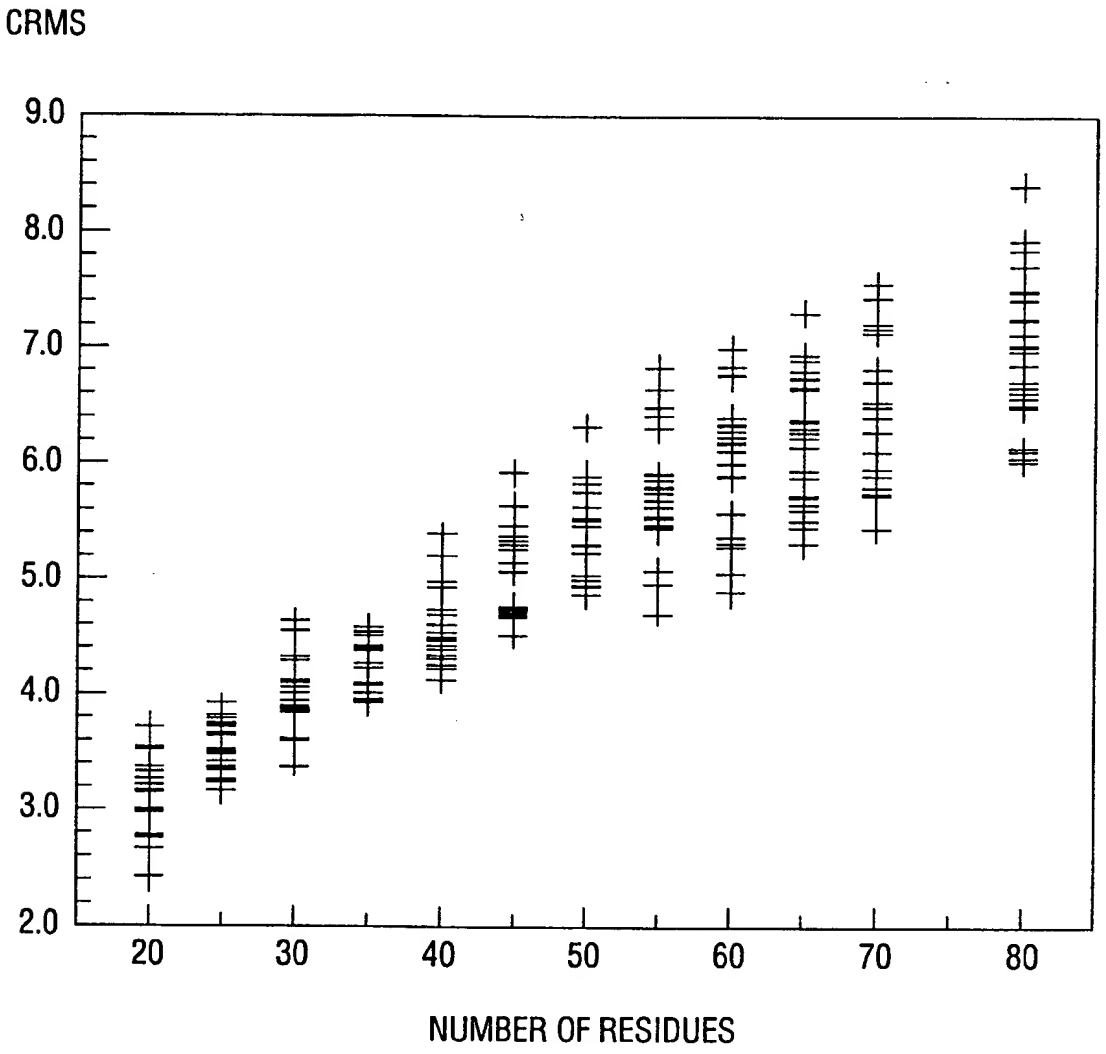

CRMS
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
20
30
40
50
60
70
80
NUMBER OF RESIDUES

FIG. 3

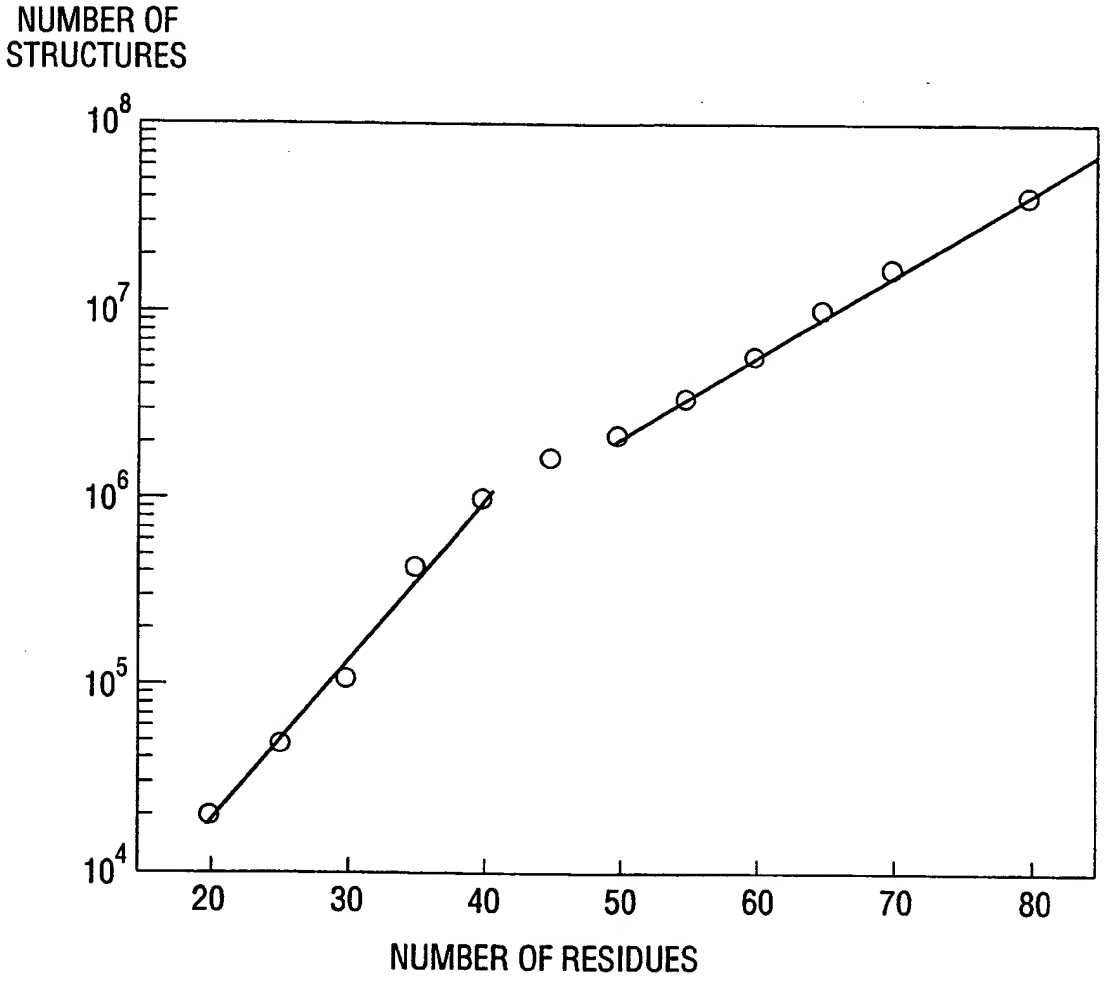

NUMBER OF STRUCTURES
10^8
10^7
10^6
10^5
10^4
20
30
40
50
60
70
80
NUMBER OF RESIDUES

FIG. 4

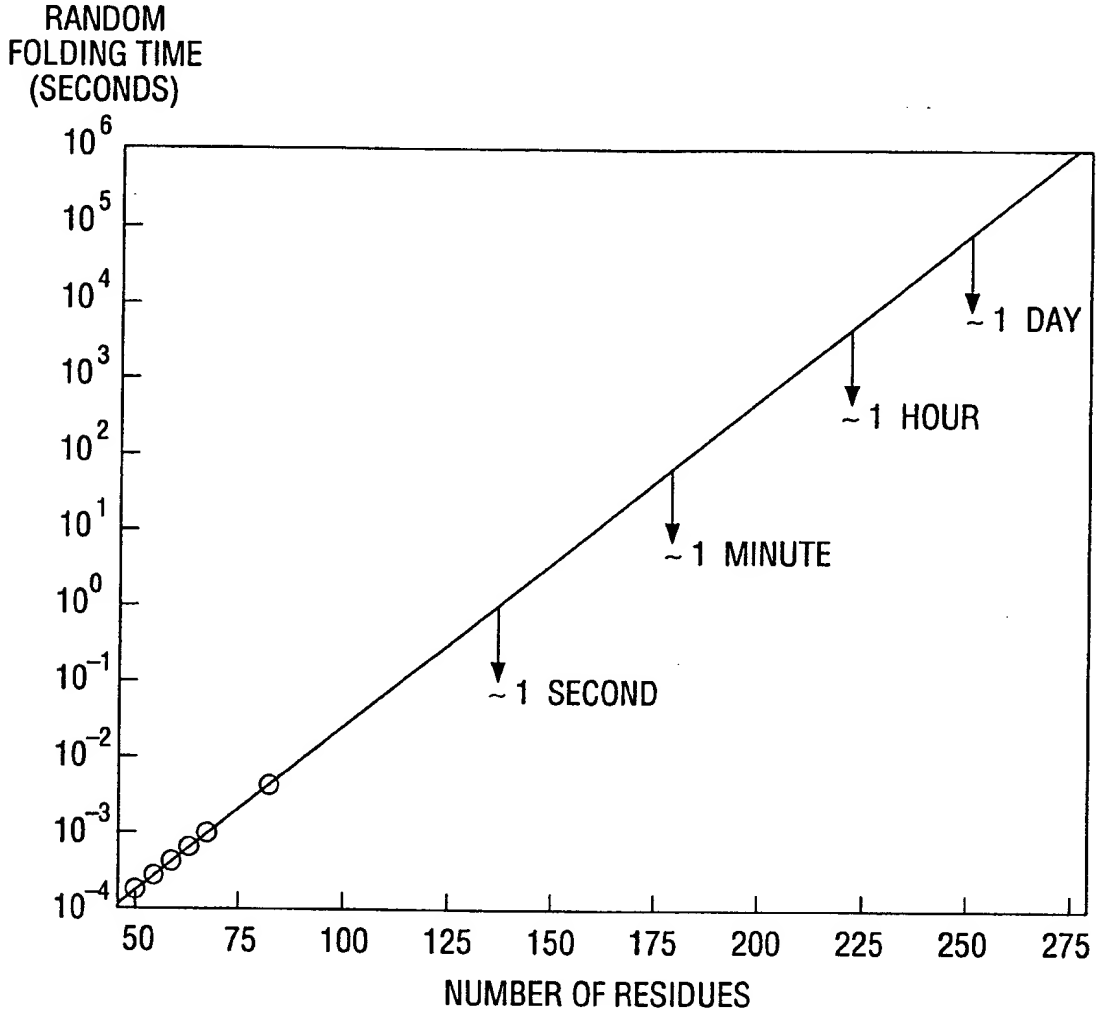

RANDOM
FOLDING TIME
(SECONDS)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
~1 DAY
~1 HOUR
~1 MINUTE
~1 SECOND
50
75
100
125
150
175
200
225
250
275
NUMBER OF RESIDUES

FIG. 5

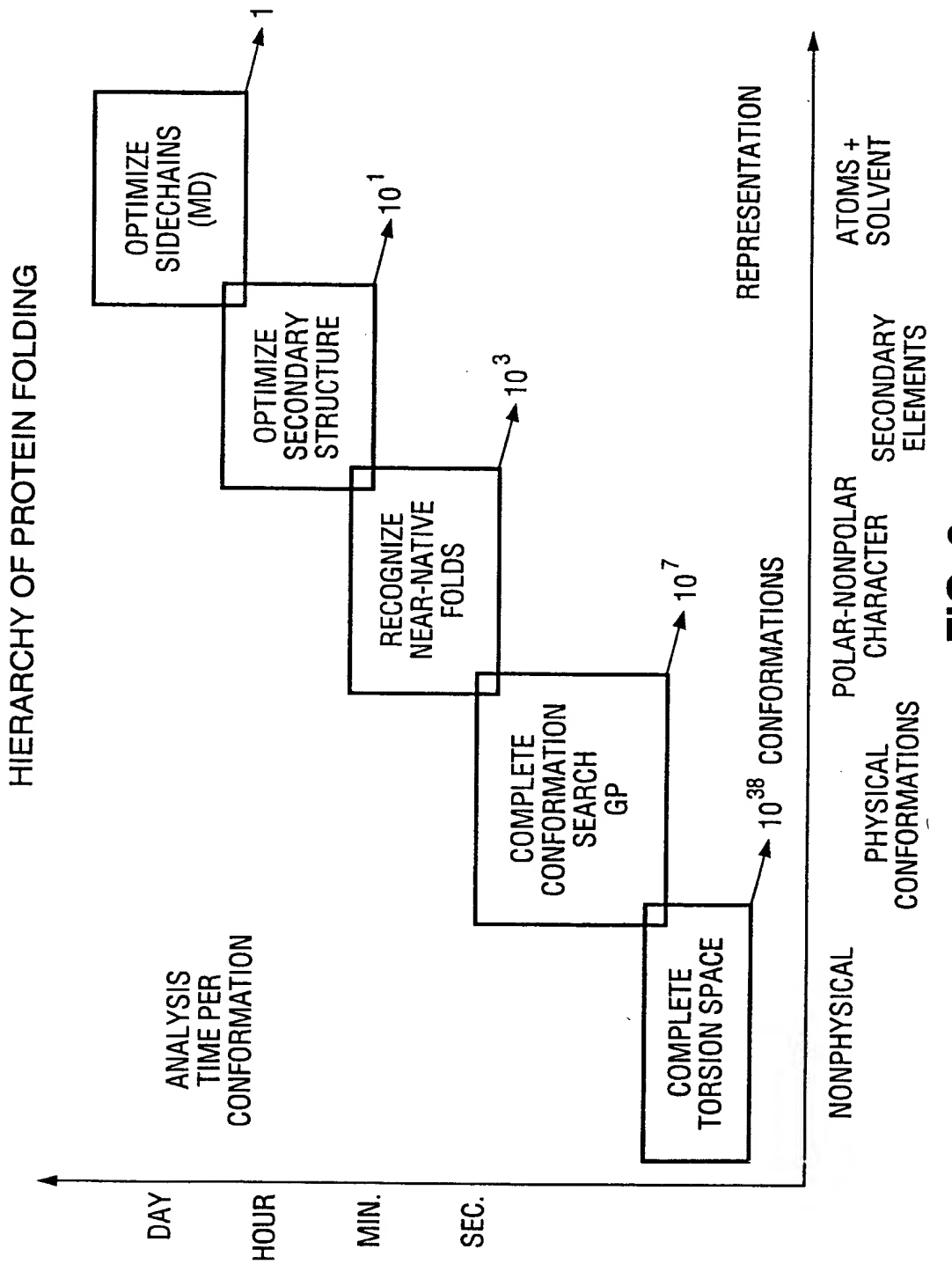

HIERARCHY OF PROTEIN FOLDING
OPTIMIZE SIDECHAINS (MD)
OPTIMIZE SECONDARY STRUCTURE
RECOGNIZE NEAR-NATIVE FOLDS
COMPLETE CONFORMATION SEARCH GP
COMPLETE TORSION SPACE
1
10^1
10^3
10^7
10^{38} CONFORMATIONS
ANALYSIS TIME PER CONFORMATION
DAY
HOUR
MIN.
SEC.
NONPHYSICAL
PHYSICAL CONFORMATIONS
REPRESENTATION
POLAR-NONPOLAR CHARACTER
SECONDARY ELEMENTS
ATOMS + SOLVENT

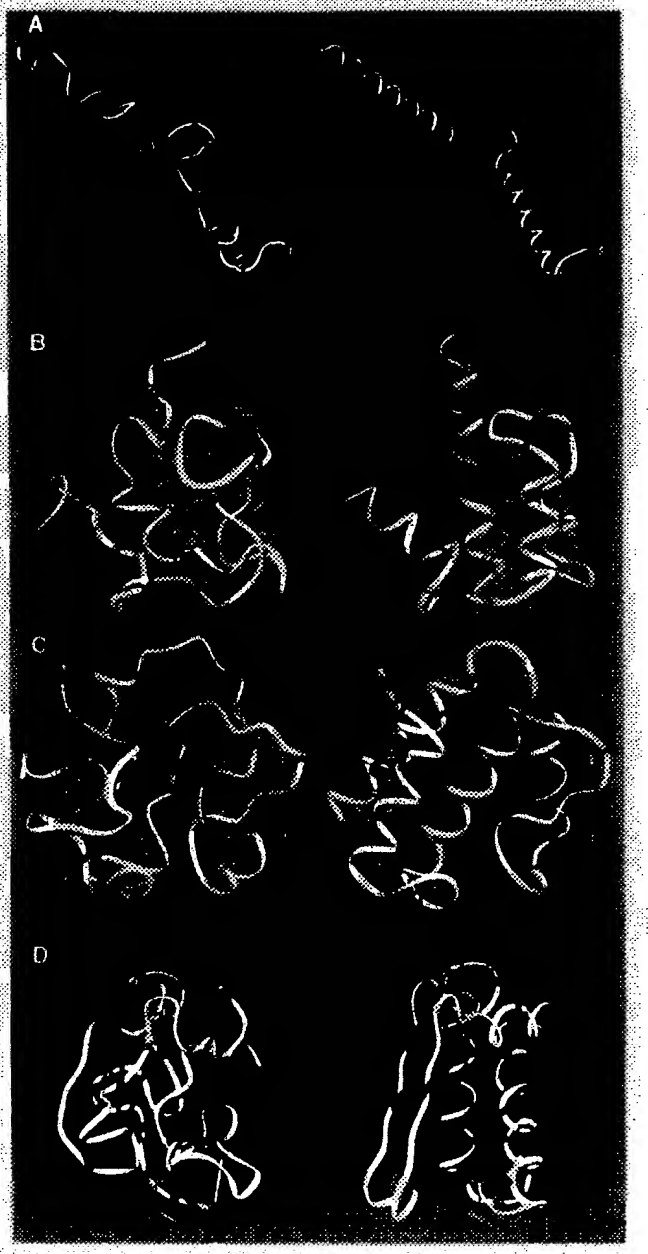
FIG. 6

FIG. 7A

FIG. 7B

FIG. 7C

FIG. 7D



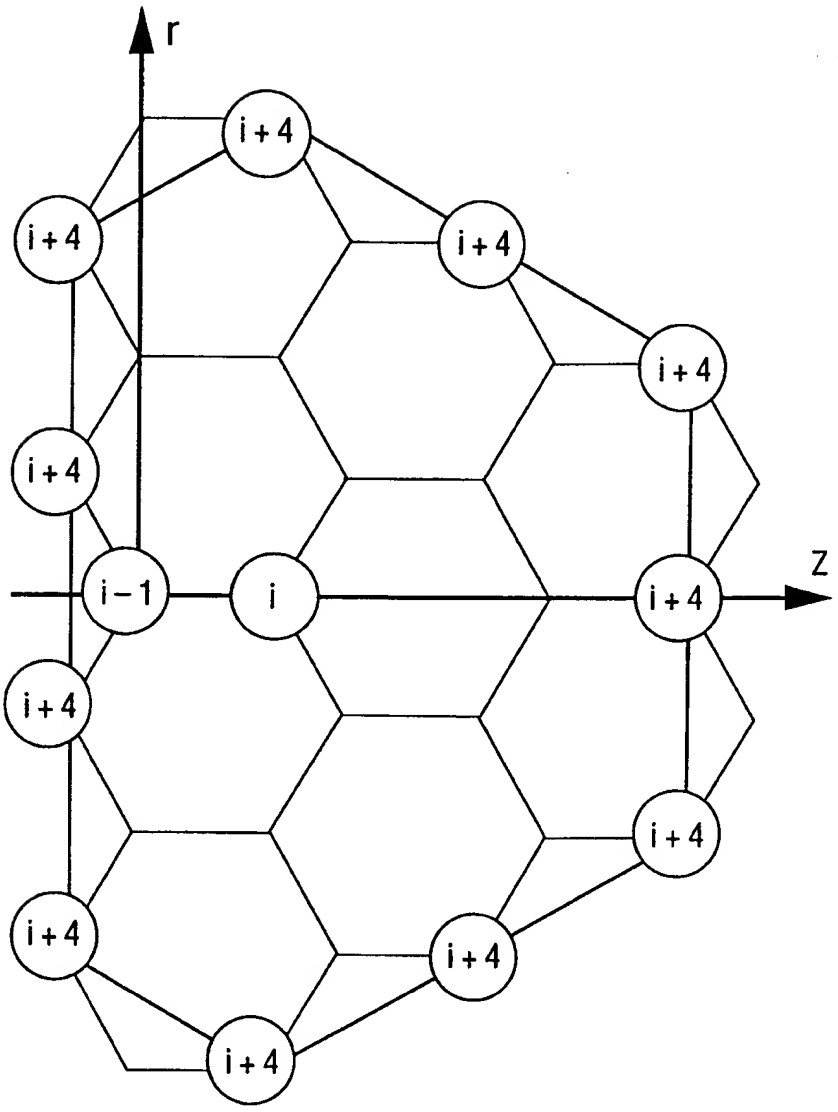


FIG. 8

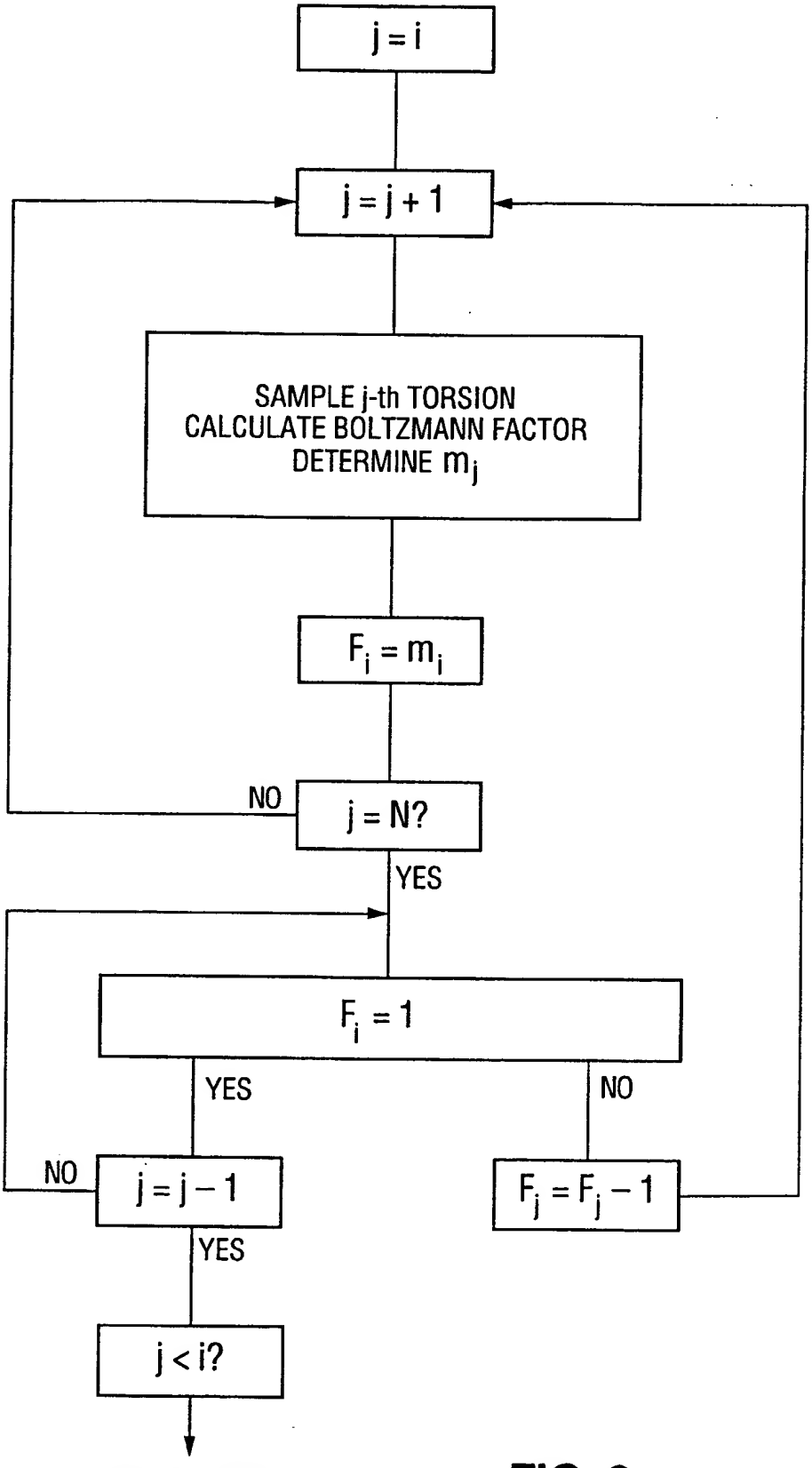

j = i
j = j + 1
SAMPLE j-th TORSION
CALCULATE BOLTZMANN FACTOR
DETERMINE m_j
$F_i = m_i$
NO
j = N?
YES
$F_i = 1$
YES
NO
NO
j = j – 1
$F_j = F_j - 1$
YES
j < i?
NEXT CLUSTER

FIG. 9

INTERNATIONAL SEARCH REPORT

Internauonal application No.
PCT/US98/08077

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G01N 33/00; G06F 17/00
US CL :364/528, 578; 436/43; 530/300; 702/27

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 364/528, 578; 436/43; 530/300; 702/27

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,241,470 A (LEE ET AL) 31 August 1993 (31/08/93), see entire document, especially column 5, lines 1-3, column 6, line 31 - column 8, line 48.	1-21
X	US 5,600,571 A (FREISNER ET AL) 04 February 1997 (04/02/97), column 4, lines 7-19, column 5, line 15 - column 6, line 39.	1-4, 7, 16
A, P	US 5,680,331 A (BLANEY ET AL) 21 October 1997 (21/10/97), see entire document.	1-21
A,P	US 5,724,252 A (IIJIMA ET AL) 03 March 1998 (03/0/3/98), see entire document.	1-21

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 25 AUGUST 1998	Date of mailing of the international search report 30 SEP 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JEFFREY E. RUSSEL Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)★

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/08077

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SCHAFER et al. Predictions of Protein Backbone Bond Distances and Angles from First Principles. Biopolymers. 1995, Volume 35, pages 603-606, especially page 603, column 2, first full paragraph.	1-4, 7, 16

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/08077

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

APS, DIALOG

search terms: protein, polypeptide, peptide, oligopeptide, angle, dihedral, torsion, backbone, initio, radius of gyration, center of mass, distance contraint